Protocolo Internacional Harmonizado para ensaios de proficiência de laboratórios analíticos (químicos)

(Relatório Técnico IUPAC)

Resumo: As organizações internacionais de normalização AOAC International, ISO, e IUPAC – cooperaram para produzir o Protocolo Internacional Harmonizado para Ensaios de Proficiência de Laboratórios Analíticos (Químicos). O Grupo de Trabalho que produziu o protocolo acordou em revisá-lo à luz dos recentes desenvolvimentos e com base na experiência obtida desde sua primeira publicação. Esta revisão foi preparada e acordada com base nos comentários recebidos a partir de consulta aberta.

Palavras-chave: harmonizado; Divisão IUPAC de Química Analítica; incerteza; análise; ensaio de proficiência; protocolo.

SUMÁRIO

PARTE 1: PREFÁCIO E INTRODUÇÃO

1.0 Prefácio

Nos 10 anos seguintes à publicação da primeira versão deste protocolo [1], os ensaios de proficiência floresceram. O método tornou-se amplamente usado em vários setores da análise química, e vários novos programas de ensaios de proficiência foram lançados mundialmente [2]. Foi implementado um estudo detalhado dos ensaios de proficiência para laboratórios de análise química [3]. A *International Organization for Standardization* (ISO) publicou um guia sobre ensaios de proficiência [4] e uma norma sobre métodos estatísticos para ensaios de proficiência [5]. A *International Laboratory Accreditation Co-operation* (ILAC) publicou um documento sobre os requisitos da qualidade para ensaios de proficiência [6], e muitos programas de ensaios de proficiência já foram, atualmente, acreditados. Além disso, o esclarecimento, ao longo da última década, sobre a aplicação do conceito de incerteza à medição química teve efeito sobre a maneira como são vistos os ensaios de proficiência. Este extraordinário desenvolvimento dos ensaios de proficiência deve-se tanto ao reconhecimento da sua incomparável capacidade de expor problemas inesperados na análise, quanto ao atual requisito de participação em programas de ensaios de proficiência como parte da acreditação de laboratórios analíticos.

Como resultado desta atividade em diferentes setores analíticos, mais o considerável número de pesquisas que foram conduzidas, a comunidade analítica acumulou um grande número de novas experiências com ensaios de proficiência. É, com satisfação, que se nota que não foram necessárias modificações substanciais nas idéias e princípios fundamentais do Protocolo Harmonizado de 1993 [1] a fim de integrar esta nova experiência. Entretanto, a experiência adicional mostrou a necessidade, e provê o fundamento, para um refinamento da abordagem sobre vários aspectos dos ensaios de proficiência, assim como de recomendações definidas e mais específicas em algumas áreas. Além disso, o Protocolo Harmonizado original tratava, na sua maior parte, da organização de programas de ensaios de proficiência e, portanto, era dirigido principalmente aos provedores de programas. A crescente importância dos dados de ensaios de proficiência gerou, entretanto, a necessidade de orientação adicional na interpretação dos resultados dos programas, tanto por parte dos participantes como dos "usuários finais" de dados analíticos (tais como clientes de laboratórios, organismos regulamentadores, e outros interessados em qualidade laboratorial). Todos estes fatores demandaram uma atualização do Protocolo Harmonizado de 1993.

A revisão também oferece uma oportunidade de mostrar que alguns aspectos importantes dos ensaios de proficiência ainda não estão completamente documentados até o momento. Também se deve reconhecer que a variedade de possíveis abordagens incluída nos documentos ISO visa ser abrangente e englobar todos os campos de medição. A experiência prática em química analítica sugere enfaticamente que um subconjunto restrito, extraído desta ampla variedade, oferece uma abordagem otimizada para o trabalho analítico rotineiro. Esta atualização do Protocolo Harmonizado não é, portanto, uma mera colagem de trechos de outros documentos, mas um subconjunto otimizado de métodos, baseado na experiência prática detalhada de condução de programas de ensaios de proficiência, interpretados especificamente para a química analítica, e incorporando novas idéias.

Publicar um Protocolo atualizado permite enfatizar a importância do julgamento profissional e experiência tanto para a execução dos programas de ensaios de proficiência quanto para a ação adequada dos participantes a partir dos resultados obtidos. A adesão a um protocolo obviamente implica que certas ações *precisam* ser implementadas. Mas os meios pelos quais elas serão implementadas precisam ser em parte pertinentes à aplicação específica, e o espectro de possíveis aplicações é amplo e variável ao longo do tempo. Ademais, qualquer analista químico experiente admitirá prontamente que, na prática, materiais e métodos de ensaios do campo de química analítica apresentarão invariavelmente um comportamento ocasionalmente inesperado, o que requer consideração e atenção dos especialistas. Portanto, entende-se que seria arriscado eliminar o julgamento profissional dos especialistas, substituindo-o por regras inflexíveis. A estrutura do presente documento reflete tal filosofia. O Protocolo em si é apresentado primeiro, e compreende uma série de seções relativamente curtas destacando as ações essenciais requeridas dos programas que afirmam aderir ao mesmo. Este documento está estruturado por várias seções e apêndices mais longos que discutem as opções disponíveis para implementar o Protocolo e as razões pelas quais são feitas recomendações específicas. Os apêndices incluem seções independentes específicas para participantes e para usuários finais de dados, a fim de auxiliá-los na interpretação dos dados dos programas de ensaios de proficiência.

Para encerrar, observe-se que embora este documento tenha o título de "Protocolo", não se está aqui defendendo a filosofia de que há apenas uma maneira correta de conduzir ensaios de proficiência. Este documento simplesmente mostra o que se verificou serem procedimentos bons e efetivos para químicos analíticos na maioria dos casos. Reconhece-se que as circunstâncias podem impor que sejam necessários procedimentos alternativos, e as escolhas destes são responsabilidade do provedor do programa, auxiliado pelo respectivo comitê assessor. Também se destaca que este protocolo restringe-se na sua maior parte aos

aspectos científicos e técnicos dos ensaios de proficiência usados como uma ferramenta para melhorar o desempenho das medições. Portanto, ele não aborda assuntos como a qualificação ou desqualificação de laboratórios com relação a determinados fins, nem a acreditação de provedores de programas de ensaios de proficiência ou programas específicos.

1.1 Os fundamentos do ensaio de proficiência

Para que um laboratório produza dados consistentemente confiáveis, deve implementar um programa adequado de procedimentos de garantia da qualidade e de monitoramento do desempenho. O ensaio de proficiência é um destes procedimentos. O formato usual dos programas de ensaios de proficiência em química analítica é baseado na distribuição de amostras de um material de ensaio para os participantes. Laboratórios participantes ("participantes") geralmente sabem que o material de ensaio foi enviado por um provedor de programa de proficiência, mas eventualmente o material pode ser recebido "às escuras" (isto é, recebido de um cliente normal do laboratório). Os participantes analisam o material sem o conhecimento do resultado correto e retornam o resultado da medição para o provedor do programa. O provedor converte os resultados em índices *(scores)* que refletem o desempenho do laboratório participante. Isto alerta o participante quanto a problemas inesperados que poderiam estar presentes, e induz a gerência a implementar a ação corretiva necessária.

A idéia principal deste Protocolo Harmonizado é que os ensaios de proficiência forneçam informações sobre a adequação aos propósitos dos resultados analíticos produzidos por seus participantes, de maneira a auxiliá-los a cumprir requisitos. Isto pode ser obtido quando:

- os critérios para avaliação de resultados levam em consideração a adequação aos propósitos, e consequentemente os índices *(scores)* informam aos participantes quando eles precisam melhorar seu desempenho para satisfazer as necessidades dos clientes (ou outros interessados);
- as circunstâncias em que foram executados os ensaios de proficiência são próximas àquelas que prevalecem durante as análises rotineiras, de forma que o resultado representa a "vida real"; e
- o método de classificação (escore) é simples, e onde possível, consistente ao longo de todo a faixa da medição analítica, assegurando sua pronta interpretação pelos participantes e clientes.

Embora o primeiro objetivo dos ensaios de proficiência seja prover a base para que cada participante ajude a si mesmo, seria falso ignorar o fato de que os resultados de ensaios de proficiência também são usados para outros fins. Os participantes normalmente usam os seus índices *(scores)* para demonstrar sua competência para clientes em potencial e avaliadores de acreditação, e isto tem como efeito negativo pressionar os analistas para obter excelência nos ensaios de proficiência e não simplesmente avaliar seus procedimentos de rotina. Os participantes devem se esforçar ao máximo no sentido de evitar tal tendência pois, normalmente, é impossível para os provedores de programa detectar ou eliminar isto. Os participantes devem também ser cuidadosos no sentido de evitar a interpretação errônea dos índices *(scores)* acumulados.

1.2 O ensaio de proficiência em relação a outros métodos de garantia da qualidade

Um programa abrangente de garantia da qualidade (GQ) em laboratórios de química analítica inclui os seguintes elementos além de ensaios de proficiência: a validação dos métodos analíticos [7]; o uso de materiais de referência certificados (MRCs) quando disponíveis [8]; e o emprego rotineiro de controle da qualidade (CQ) interno [9]. Tradicionalmente, a validação de um método analítico implica que suas características de desempenho – veracidade, precisão sob condições variadas, linearidade de calibração, e assim por diante—são suficientemente bem conhecidas. Em termos atuais, significa que a incerteza de medição do método é estimada numa operação única sob condições normais de uso, verificando-se que ele é potencialmente adequado para a finalidade. Idealmente, a validação de métodos envolve, entre outras coisas, o uso de MRCs de matriz adequada, se disponíveis, para calibração ou para verificação de calibrações existentes, caso efeitos matriz estejam presentes. Quando MRCs não são disponíveis, outros expedientes têm que ser empregados.

Convém que o CQ interno seja conduzido como rotina e envolva a análise de um ou mais "materiais de controle" dentro de cada corrida da análise. Este procedimento, combinado com o uso de gráficos de controle, assegura que os fatores determinantes da magnitude da incerteza não tenham se modificado substancialmente desde que a adequação aos propósitos foi originalmente demonstrada no processo de validação. Em outras palavras, a incerteza estimada na validação é demonstrada (dentro dos limites de variação estatística) para aplicação em cada corrida individual executada subseqüentemente. A preparação de um material de controle também envolve idealmente o uso de MRCs, a fim de se estabelecer a rastreabilidade dos valores atribuídos ao mensurando.

Em princípio, a validação de métodos e o CQ interno são suficientes para assegurar exatidão. Na prática, no entanto, eles freqüentemente não são perfeitos. O ensaio de proficiência é, portanto, o meio de assegurar

que estes dois procedimentos intralaboratoriais funcionem satisfatoriamente. Na validação de métodos, influências desconhecidas podem interferir no processo de medição e, em muitos setores, MRCs não estão disponíveis. Sob tais condições, a rastreabilidade é difícil de se estabelecer, e fontes de erro não identificadas podem estar presentes no processo de medição. Laboratórios sem uma referência externa podem operar por longos períodos com variações aleatórias ou tendências de magnitude significativas. O ensaio de proficiência é um meio de detectar e iniciar a correção de tais problemas (ver Apêndice 6). Sua principal vantagem consiste em prover um meio através do qual os participantes podem obter uma avaliação externa e independente da exatidão de seus resultados.

PARTE 2: O PROTOCOLO HARMONIZADO; ORGANIZAÇÃO DE PROGRAMAS DE ENSAIOS DE PROFICIÊNCIA

2.1 Objetivo e campo de aplicação

Este protocolo é aplicável onde:

- o objetivo principal é a avaliação do desempenho de um laboratório através de critérios estabelecidos com base na adequação a um propósito comum;
- a conformidade a estes critérios pode ser julgada com base no desvio dos resultados de medição emrelação aos valores designados; e
- os resultados dos participantes são relatados num intervalo de escala ou de razão.

Nota: Estas condições são amplamente aplicáveis na avaliação de execução de ensaios de rotina por laboratórios de análise química, mas também em muitos outros campos de medição e ensaio.

Este protocolo não se destina à avaliação de serviços de calibração e portanto não se destina ao uso, pelo provedor do programa, de dados sobre incerteza agregados aos resultados dos participantes. Tampouco provê critérios para a avaliação, certificação ou acreditação de provedores de programas de ensaios de proficiência.

2.2 Terminologia

Neste documento, são utilizadas as definições da ISO , quando disponíveis. As abreviações obedecem ao Compêndio de Nomenclatura Analítica IUPAC (1997). Os seguintes termos adicionais são utilizados com freqüência neste documento:

- *Provedor de programa de ensaio proficiência ("o provedor do programa" ou "provedor"):* Organização responsável pela coordenação de um programa de ensaio de proficiência específico.
- *Material de ensaio (de proficiência):* material que é distribuído para ensaio pelos participantes num programa de ensaios de proficiência.
- *Unidade de distribuição:* porção embalada de material de ensaio que é enviada ou está pronta para ser enviada a um laboratório participante.
- *Porção de ensaio:* parte da unidade de distribuição que é usada para análise.
  Nota: Uma porção de ensaio pode consistir em toda uma unidade de distribuição ou uma porção da mesma.
- *Série:* Parte de um programa de ensaio de proficiência, definida por uma série de materiais de ensaio, analitos, métodos analíticos, ou outras características comuns.
- *Rodada:* Um evento individual de distribuição dentro de uma série.

2.3 Estrutura do ensaio de proficiência

2.3.1 Operação do programa

- Os materiais de ensaio devem ser distribuídos periodicamente aos participantes, os quais devem retornar resultados dentro de um prazo determinado.
- Um valor é designado para cada mensurando, antes ou depois da distribuição; este valor não é informado aos participantes até o final do prazo de entrega dos resultados.
- Os resultados são submetidos à análise estatística e/ou convertidos em índices *(scores)* pelo provedor do programa, e os participantes devem ser imediatamente informados sobre seus desempenhos.
- Orientação deve ser disponibilizada para os participantes com desempenho fraco, e todos os participantes devem ser inteiramente informados sobre o andamento do programa.

2.3.2 A estrutura de uma dada rodada

Convém que a estrutura do programa, para qualquer analito ou rodada dentro de uma série, seja a seguinte:

- o provedor do programa organiza a preparação e validação do material de ensaio;
- o provedor do programa distribui as unidades do material de ensaio de acordo com o cronograma;
- os participantes analisam os materiais de ensaio e relatam os resultados ao provedor;
- os resultados são submetidos à análise estatística e/ou atribuição de índices *(scores)*;
- os participantes são informados sobre seus desempenhos;
- quando solicitado, o provedor do programa, disponibiliza orientação aos participantes com desempenho fraco; e
- o provedor do programa deve efetuar analise critica do desempenho do programa durante uma rodada particular, e implementa os ajustes necessários.

Nota: Freqüentemente, a preparação para uma rodada do programa terá que ser organizada enquanto a rodada anterior está em andamento.

2.4 Organização

- A operação rotineira do programa é responsabilidade do provedor do programa.
- O provedor do programa deve documentar todas práticas e procedimentos no seu próprio manual da qualidade, e deve fornecer a todos os participantes um resumo dos procedimentos relevantes.
- Convém também que o provedor do programa mantenha todos os participantes informados sobre a eficácia do programa como um todo, sobre quaisquer mudanças introduzidas, e sobre como têm sido tratados os problemas encontrados.
- A operação do programa deve passar periodicamente por uma análise crítica (ver abaixo).
- A condução geral do programa deve ser monitorada por um comitê assessor com representantes (deve incluir entre eles químicos analíticos com experiência no campo específico), por exemplo, do provedor do programa, dos laboratórios contratados (se houver), dos órgãos profissionais relacionados, dos participantes, e dos usuários finais dos dados analíticos. O comitê assessor deve também incluir um especialista em estatística.

2.5. Responsabilidades do comitê assessor

O comitê assessor considerará e oferecerá orientação sobre os seguintes tópicos:

- a escolha dos tipos de materiais de ensaio, analitos e as faixas de concentração dos analitos
- a freqüência das rodadas
- o sistema de índice *(score)* e os procedimentos estatísticos (incluindo aqueles usados em ensaios de homogeneidade)
- a orientação que pode ser oferecida aos participantes
- problemas gerais e específicos que surjam durante a operação do programa
- as instruções enviadas aos participantes
- o formato dos relatórios de resultados preenchidos pelos participantes
- o conteúdo dos relatórios enviados aos participantes
- outros meios de comunicação com os participantes
- os comentários dos participantes e usuários finais com relação à operação do programa
- o nível de confidencialidade adequado ao programa

2.6 Análise crítica do programa

- A operação do programa deve sofrer uma análise crítica periódica.
- O provedor do programa deverá analisar criticamente os resultados de cada rodada do programa, observando, por exemplo, quaisquer pontos fortes e pontos fracos, problemas específicos, e oportunidades de melhoria.
- O provedor e o comitê assessor devem considerar cada aspecto da operação do programa, incluindo os pontos identificados pela análise crítica do provedor do programa sobre cada rodada, normalmente em intervalos de um ano.
- Um resumo desta análise crítica deve estar disponível aos participantes e outros, conforme adequado e acordado pelo comitê assessor.

2.7 Materiais de ensaio

- O provedor do programa deve providenciar a preparação dos materiais de ensaio. A preparação dos materiais de ensaio e outros aspectos do programa podem ser subcontratados, mas o provedor continua sendo o responsável, devendo exercer o devido controle.
- Convém que a organização que preparar o material de ensaio tenha experiência demonstrável na área de análise que está sendo ensaiada.
- Os materiais de ensaio a serem distribuídos no programa devem geralmente ser similares, quanto ao tipo, aos materiais que são rotineiramente analisados (no que tange à composição da matriz e a faixa de concentração, quantidade, ou nível do analito).
- O lote de material preparado para o ensaio de proficiência deve ser suficientemente homogêneo e estável, no que concerne a cada analito, de forma a assegurar que todos os laboratórios recebam unidades de distribuição que não se diferenciem em qualquer grau relevante quanto à concentração de analito média (ver Seção 3.11).
- O provedor do programa deve declarar claramente o procedimento usado para determinar a homogeneidade do material de ensaio.

Nota: Se por um lado é exigido que a homogeneidade entre as unidades seja suficiente, por outro lado não convém que o participante suponha que a sua unidade de distribuição é suficientemente homogênea em si mesma para o seu procedimento analítico específico. É responsabilidade dos participantes assegurar que a porção de ensaio usada para análise seja representativa do material de ensaio como um todo, contido na unidade de distribuição.

- A quantidade de material numa unidade de distribuição deve ser suficiente para a análise requerida, incluindo qualquer reanálise, quando permitido pelo protocolo do programa.
- Ao se avaliar analitos instáveis, pode ser necessário que o provedor do programa prescreva uma data limite até a qual a análise deva ser completada.
- Os provedores de programas devem considerar quaisquer riscos que os materiais de ensaio possam oferecer e tomar as medidas adequadas para avisar toda e qualquer parte envolvida (exemplo: distribuidores de materiais de ensaio, laboratórios de ensaios, etc.) sobre o risco potencial inerente.

  Nota: As "medidas adequadas" incluem, mas não se limitam, à conformidade com a legislação específica. Muitos países também impõem um "dever de cuidar" adicional, o qual pode se estender além dos requisitos mínimos legais.

- Os participantes devem receber, juntamente com os materiais de ensaio, informação suficiente sobre os materiais, e quaisquer critérios de adequação aos propósitos que serão aplicados, a fim de lhes permitir selecionar os métodos adequados de análise. Esta informação não deve incluir o valor designado.

2.8 Freqüência de distribuição

A freqüência de distribuição adequada para os materiais de ensaio deverá ser decidida pelo provedor do programa com auxílio do comitê assessor (ver Seção 3.10). Normalmente é de 2 a 10 rodadas por ano.

2.9 Valor designado

O valor designado é uma estimativa do valor do mensurando que é usada para fins de calcular índices *(scores)*.

- O valor designado deverá ser determinado por um dos seguintes métodos:
  - determinação por um laboratório de referência [1]
  - o(s) valor(es) certificado (s) de um MRC usado como material de ensaio
  - comparação direta do material de ensaio de proficiência com os valores de consenso de MRCs de laboratórios especialistas
  - formulação (ou seja, designação de valor com base nas proporções usadas numa solução ou outra mistura de ingredientes com conteúdo conhecido de analito)

[1] Um "laboratório de referência", neste contexto, consiste num laboratório consensualmente visto pelo provedor do programa e pelo comitê assessor como provedor de valores de referência de suficiente confiabilidade para os fins do programa.

- um valor de consenso (isto é, um valor derivado diretamente dos resultados relatados)

O valor designado só será informado aos participantes depois da data estabelecida para entrega dos resultados.

- O provedor de programa deve relatar o valor designado e uma estimativa de sua incerteza aos participantes quando relatar os resultados e índices *(scores)*, dando detalhes suficientes sobre como o valor designado e incerteza foram determinados. Métodos para determinação do valor designado são discutidos abaixo (ver Seção 3.2).
- Naqueles setores onde são empregados métodos empíricos de análise, convém que o valor designado seja normalmente calculado a partir dos resultados obtidos através do uso de um procedimento analítico claramente definido. Alternativamente, o valor designado pode ser calculado a partir dos resultados de dois ou mais métodos empíricos que demonstrem ser efetivamente equivalentes.
- Pode ser eventualmente necessário que o programa use valores designados diferentes para diferentes métodos, mas convém que esse expediente seja usado apenas para cumprir uma necessidade específica.
- Quando o valor designado advém de método empírico, os participantes devem ser notificados previamente qual procedimento empírico será usado para a determinação do valor designado.

2.10 Escolha do método analítico pelo participante

- Os participantes normalmente deverão usar o método analítico de sua escolha. Em alguns casos, entretanto, por exemplo, quando a legislação assim o exigir, os participantes pode ser instruídos a usar um método especificamente documentado.
- Os métodos devem ser aqueles usados pelo participante em trabalho de rotina no setor apropriado, e não versões do método especialmente adaptadas para o ensaio de proficiência.

2.11 Avaliação de desempenho

Os laboratórios serão avaliados com base na diferença entre o seu resultado e o valor designado. Um índice *(score)* de desempenho será calculado para cada laboratório, usando o esquema estatístico detalhado na Seção 3.1.

Nota: O índice-z (z-score) (z-score) baseado num critério de adequação ao propósito é o único tipo de índice *(score)* recomendado neste protocolo.

2.12 Critérios de desempenho

Para cada analito numa rodada deve ser estabelecido um critério de desempenho, contra o qual poderá ser julgado o desempenho obtido pelo laboratório. O critério de desempenho será estabelecido de maneira a assegurar que os dados analíticos rotineiramente produzidos pelo laboratório sejam de uma qualidade que é adequada aos propósitos perseguidos. Ele não será estabelecido para representar o melhor desempenho que métodos típicos são capazes de prover (ver Seção 3.5).

2.13 Relatório de resultados pelos participantes

- Os participantes devem relatar os resultados usando o método e formato exigidos pelo programa.
- O provedor do programa deve estabelecer a data até a qual os resultados devem ser relatados. Resultados apresentados depois do prazo deverão ser rejeitados.
- Os resultados apresentados não podem ser corrigidos ou retirados.

Nota: O motivo para esta abordagem rigorosa é que o ensaio de proficiência existe para ensaiar todos aspectos de obter e produzir um resultado analítico, incluindo calcular, verificar e relatar um resultado.

2.14 Relatórios fornecidos pelo provedor do programa

- O provedor do programa deve fornecer um relatório de desempenho para cada participante para cada rodada.
- Os relatórios emitidos para os participantes devem ser claros e abrangentes e mostrar a distribuição dos resultados de todos os laboratórios, juntamente com o índice *(score)* de desempenho do participante.

- Convém que os resultados de ensaios usados pelo provedor do programa também estejam disponíveis, para permitir aos participantes verificar se os seus dados foram corretamente informados.
- Os relatórios devem ser colocados à disposição tão rápido quanto possível depois da entrega dos resultados ao laboratório coordenador e, se possível, antes da próxima distribuição de amostras.
- Os participantes devem receber, pelo menos: (a) relatórios num formato simples e claro, e (b) resultados de todos os laboratórios em formato gráfico (ex: histograma, gráfico de barras, ou outros gráficos de distribuição) com o apropriado sumário estatístico.

Nota: Embora seja recomendável que todos os resultados sejam relatados aos participantes, pode não ser possível alcançar esse objetivo em alguns programas muito grandes (ex: onde há centenas de participantes, cada um determinando 20 analitos em uma dada rodada).

2.15 Relação com os participantes

- Ao aderir ao programa, os participantes devem receber informação detalhada descrevendo:

- a gama de ensaios disponíveis e os ensaios que o participante escolheu para implementar;
- o método para se estabelecer os critérios de desempenho;
- os critérios de desempenho aplicáveis no momento da adesão, a não ser que sejam estabelecidos critérios em separado para cada material de ensaio;
- o método de determinação dos valores designados, incluindo métodos de medição onde pertinente;
- um resumo dos procedimentos estatísticos usados para se obter os índices *(scores)* dos participantes;
- informação sobre a interpretação de índices *(scores)*;
- os requisitos relativos à participação (ex: pontualidade nos relatórios, como evitar a colusão com outros participantes, etc);
- a composição e o método de seleção do comitê assessor; e
- os dados para contato com o provedor e qualquer outra organização pertinente.

Nota: A comunicação com os participantes pode se dar através de qualquer meio apropriado, incluindo, por exemplo, os boletins periódicos, o relatório habitual de análise critica do programa, as reuniões abertas periódicas, ou comunicação eletrônica.

- Os participantes devem ser avisados sobre quaisquer mudanças planejadas no projeto ou operação do programa.
- Deve haver orientação para participantes com desempenho insatisfatório, embora isto possa acontecer na forma de uma lista de consultores especialistas no campo específico.
- Os participantes que considerem haver erro na avaliação do seu desempenho devem poder submeter a questão ao provedor do programa.
- Deve haver um mecanismo através do qual os participantes possam comentar aspectos da operação do programa e problemas com materiais de ensaio específicos de forma a contribuir para o desenvolvimento do programa e permitir aos participantes alertar o provedor do programa sobre qualquer dificuldade não prevista com materiais de ensaio.
  Nota: É recomendável que o *feedback* dos participantes seja incentivado.

2.16 Colusão e falsificação de resultados

- É responsabilidade dos laboratórios participantes evitar colusão ou falsificação de resultados. Esta deve ser uma condição escrita de participação num programa, incluída nas instruções aos participantes do programa.
- O provedor do programa deverá envidar os devidos esforços no sentido de desestimular a colusão, através da concepção adequada do programa. (Por exemplo, pode ser divulgado que mais de um material de ensaio pode ser eventualmente distribuído durante uma rodada, de forma que os laboratórios não possam comparar os resultados diretamente, sendo recomendável que não haja reutilização identificável de materiais em rodadas consecutivas.)

Nota: A colusão, seja entre participantes ou entre participantes individuais e o provedor do programa, é contrária à conduta cientifica profissional e serve apenas para anular os benefícios dos ensaios de proficiência para os clientes, organismos de acreditação, e analistas da mesma forma. A colusão deve, portanto, ser energicamente desestimulada.

2.17 Repetitividade

Convém que a média das determinações replicadas de amostras de ensaios de proficiência seja relatada apenas se este for o procedimento de rotina do laboratório. (É recomendável que os procedimentos usados pelos laboratórios participantes de programas de ensaios de proficiência simulem aqueles usados na rotina do laboratório)

Nota: O relato em separado dos resultados de determinações replicadas de um laboratório é possível em ensaios de proficiência, mas não é recomendado. Se a prática for adotada, os provedores e os participantes do programa devem se acautelar para não interpretar erroneamente os desvios-padrão da repetitividade como a média de muitos participantes. Por exemplo, a soma dos quadrados intragrupo obtida por análise da variância não pode ser interpretada como uma variância de repetitividade "média" quando diferentes métodos analíticos estão sendo usados -.

2.18 Confidencialidade

O grau de confidencialidade adotado pelo provedor do programa e pelos participantes referente às informações do programa e aos dados emitidos pelos participantes, deve estar estabelecido nas condições de participação e ser notificado aos participantes antes de sua adesão ao programa.

Nota: Ao estabelecer as condições de confidencialidade, convém que os organizadores considerem o benefício da ampla disponibilidade dos dados gerais de desempenho para a comunidade analítica e, portanto, são eles aqui encorajados a oferecer para livre publicação tais informações, observando-se a devida proteção aos dados sobre participantes individuais.

A não ser que de outra forma disposto nas condições de participação:

- O provedor do programa não deve revelar a identidade de um participante a terceiros, inclusive a outros participantes, sem o consentimento expresso do participante em questão.
- Os participantes deverão ser identificados nos relatórios apenas por um código.

Nota: A designação aleatória de códigos de laboratórios para cada rodada evita a identificação com base no histórico de participação, e é recomendada onde praticável.

- Os participantes podem comunicar seus próprios resultados, incluindo os relatórios habituais do programa, de forma privada a um organismo de acreditação de laboratórios ou outro organismo de avaliação, quando requerido para fins de avaliação, ou para clientes (incluindo a organização a que se subordina, se aplicável) para fins de demonstrar capacidade analítica.
- Os participantes podem publicar informações sobre seu próprio desempenho, mas não publicarão informações comparativas com outros participantes, incluindo a ordenação de índices *(scores)*.

PARTE 3: IMPLEMENTAÇÃO PRÁTICA

3.1 Conversão dos resultados dos participantes em índices *(scores)*

3.1.1 Os fundamentos do conceito de índice (score)

O Protocolo Harmonizado de 1993 recomendava a conversão dos resultados dos participantes em índices *(scores)* de índices-z (*z-scores*), e a experiência nos anos seguintes demonstrou a ampla aplicabilidade e aceitação do índice-z (*z-score*) em ensaios de proficiência. O resultado x de um participante é convertido num índice-z (*z-score*) de acordo com a equação:

$$z = (x - x_a) / s_p \qquad (1)$$

onde x_a é o "valor designado", a melhor estimativa do provedor do programa sobre o valor do mensurando (o valor verdadeiro da concentração do analito presente no material do ensaio de proficiência) e s_p éo

"desvio-padrão para o ensaio de proficiência em questão" baseado na adequação ao propósito. Orientações sobre a estimativa de x_a e σ_p são dadas abaixo (ver Seções 3.2 – 3.5).

Nota 1: σ_p foi definido como "valor alvo" no Protocolo Harmonizado de 1993 [1]. Considera-se hoje que tal terminologia pode levar a enganos.

Nota 2: No ABNT ISO Guia 43 [4] e na ISO 13528 [5], o símbolo $\hat{\sigma}$ é usado como o desvio-padrão para um dado ensaio de proficiência. σ_p é usado aqui para enfatizar a importância de atribuir uma amplitude apropriada a um propósito específico.

A idéia principal do índice-z (*z-score*) é tornar todos os índices *(scores)* dos ensaios de proficiência comparáveis, de forma que a significância de um índice *(score)* possa ser facilmente identificada, não importando a concentração ou identidade do analito, a natureza do material de ensaio, o princípio físico que fundamenta a medição analítica, ou o provedor do programa. Idealmente, convém que um índice *(score)* de, digamos z = -3,5 , independentemente de sua origem, tenha o mesmo efeito imediato para qualquer um: provedor, participante ou usuário final envolvido com ensaios de proficiência. Este requisito está intimamente ligado à idéia de adequação ao propósito). Na equação que define z, o termo $(x - x_a)$ representa o erro na medição. O parâmetro σ_p descreve a incerteza-padrão que é mais apropriada para a área de aplicação dos resultados da análise, ou em outras palavras, "adequada ao propósito". Ela não é necessariamente próxima da incerteza associada aos resultados relatados. Assim, embora se possa interpretar o índice-z (*z-score*) com base na distribuição de Gauss reduzida, não se pode esperar que estejam de acordo com aquela distribuição.

A incerteza que é adequada aos propósitos num resultado de medição depende da aplicação. Por exemplo, enquanto uma incerteza-padrão relativa [isto é, $u(x)/x$] de 10% é provavelmente adequada para muitas medições ambientais, uma incerteza relativa muito menor é requerida para ensaiar um carregamento de sucata contendo ouro a fim de determinar o seu valor comercial. Porém, existe mais além disso. Definir-se a incerteza adequada ao propósito é uma escolha equilibrada entre os custos da análise e os custos de tomar decisões incorretas. A obtenção de incertezas menores requer gastos desproporcionalmente maiores com análise. Mas empregar métodos com incertezas maiores significa uma possibilidade maior de tomar uma decisão incorreta e dispendiosa com base nos dados. A adequação ao propósito é definida como a incerteza que equilibra estes fatores, isto é, que minimiza a perda total esperada [10]. Os analistas e seus clientes geralmente não fazem uma análise matemática formal da situação, mas convém que pelo menos concordem sobre o que abrange a adequação ao propósito para cada aplicação específica.

3.1.2 Como interpretar os índices-z (z-scores)?

É importante enfatizar que a interpretação dos índices-z (*z-scores*) não é, geralmente, baseada em resumos estatísticos que descrevem os resultados observados dos participantes. Ao invés disso, usa-se um modelo adotado com base no critério de adequação ao propósito definido pelo provedor do programa, o qual é representado pelo desvio-padrão para avaliação da proficiência σ_p . Mais especificamente, a interpretação é baseada na distribuição de Gauss (também conhecida como normal) $x \sim N(x_{true}, \sigma_p^2)$, onde x_{true} é o valor verdadeiro para a grandeza que está sendo medida. A partir deste modelo, e assumindo que o valor designado está bem próximo de x_{true} de forma que os índices-z (*z-scores*) seguem a distribuição de Gauss padronizada:

- Um índice *(score)* de zero significa um resultado perfeito. Isto raramente acontecerá, mesmo nos mais competentes laboratórios.
- Aproximadamente 95 % dos índices-z (*z-scores*) se situarão entre -2 e +2. O sinal (i.e., - ou +) do índice *(score)* indica um erro negativo ou positivo respectivamente. Índices *(scores)* nesta faixa são comumente designados como "aceitável" ou "satisfatório".
- Um índice *(score)* fora da faixa de -3 a 3 seria muito incomum, indicando que convém que a causa do evento seja investigada e remediada. Índices *(scores)* nesta classe são comumente designados como "inaceitável" ou "insatisfatório", embora seja preferível uma frase não pejorativa tal como "requer ação".
- Índices *(scores)* nas faixas de -2 a -3 e 2 a 3 seriam esperados 1 vez em 20, de forma que um evento

isolado deste tipo não tem um grande significado. Índices *(scores)* nesta classe são comumente designados como "questionável".

Poucos laboratórios, talvez nenhum, enquadram-se exatamente no disposto acima. A maioria dos participantes operarão com uma média tendenciosa e com um desvio-padrão a cada rodada que difere de σ_p. Alguns gerarão resultados extremos devido a erros grosseiros. Entretanto, o modelo serve como um guia adequado para ações baseadas nos índices-z (*z-scores*) recebidos por todos os participantes, pelas seguintes razões. Uma média tendenciosa ou um desvio-padrão maior que σ_p sempre produzirá, ao longo da rodada, uma proporção maior de resultados com $|z| > 2$ e $|z| > 3$ do que o modelo de Gauss padronizado (isto é, aproximadamente 0,05 e 0,003, respectivamente). Isto alertará corretamente o participante sobre o problema. Inversamente, o participante com uma média não tendenciosa e desvio-padrão igual ou menor que σ_p produzirá uma pequena proporção de tais resultados, e receberá corretamente menos relatórios adversos.

3.2 Métodos para determinação do valor designado

Há várias abordagens possíveis que o provedor dos ensaios de proficiência pode empregar para determinar o valor designado e sua incerteza. Todas apresentam pontos fortes e pontos fracos. A seleção do método adequado para sua determinação em programas diferentes, ou mesmo em rodadas diferentes dentro de um programa ou série, dependerá, portanto, dos propósitos do programa. Ao selecionar os métodos de para determinação destes valores, convém que os organizadores de programas e grupo consultivo considerem o seguinte:

- os custos para o organizador e participantes — custos altos podem desestimular a participação e assim reduzir a efetividade do programa.
- quaisquer requisitos legais de consistência em relação a laboratórios de referência ou outras organizações.
- a necessidade de valores designados independentes, de forma a prover uma verificação de tendência para a população como um todo.
- quaisquer requisitos específicos de rastreabilidade de um dado valor de referência .

Nota: Espera-se a implementação da rastreabilidade metrológica pelos participantes como um elemento essencial de uma boa garantia da qualidade (GQ). Quando a rastreabilidade metrológica e métodos adequados de garantia da qualidade/controle da qualidade (GQ/CQ) — particularmente a validação com uso de MRCs de matriz apropriada—são devidamente implementados, um bom consenso e uma baixa dispersão dos resultados são esperados. A simples observação da dispersão dos resultados (e, particularmente, a pequena fração de laboratórios alcançando índices *(scores)* aceitáveis) já representa um teste direto da rastreabilidade efetiva, independentemente do valor designado. Entretanto, ensaiar contra um valor rastreável designado independentemente pode prover uma verificação adicional útil da rastreabilidade efetiva.

3.2.1 Medição por um laboratório de referência

Em princípio, um valor designado e uma incerteza podem ser obtidos por um laboratório de medições devidamente qualificado, usando um método com incerteza suficientemente pequena. Para a maioria dos fins práticos, isto equivale exatamente ao uso de um MRC (abaixo). Isto é vantajoso no sentido de que o material é efetivamente adaptado aos requisitos do programa. A principal desvantagem é que isto pode requerer esforço e custos desproporcionais se, por exemplo, investigações substanciais são requeridas para validar a metodologia para o material em questão ou para eliminar a possibilidade de interferências significativos.

3.2.2 Uso de um material de referência certificado

Se um MRC está disponível em quantidade suficiente para uso em um ensaio de proficiência, o(s) valor(es) certificado(s) e incerteza(s) associada(s) podem ser usados diretamente. Isto é rápido e simples de se implementar, e (geralmente) fornece um valor independente dos resultados dos participantes. A rastreabilidade adequada para o valor de referência também é automaticamente fornecida (por definição). Contudo, há desvantagens. Os MRCs de matriz natural não estão geralmente disponíveis em quantidade suficiente e/ou a custo adequado para serem usados regularmente em ensaios de proficiência. Eles podem ser facilmente identificados pelos participantes, os quais poderiam então inferir o seu valor certificado.

Finalmente, embora os ensaios de proficiência sejam geralmente de grande valia, são mais úteis nos setores analíticos onde os materiais de referência são escassos ou indisponíveis.

3.2.3 Comparação direta do material de ensaio de proficiência com materiais de referência certificados

Neste método, o material de ensaio é analisado várias vezes em paralelo com MRCs apropriados, em ordem aleatória sob condições de repetitividade (isto é, numa rodada individual) por um método com incerteza adequadamente pequena. Ficando assegurado que os MRCs são intimamente comparáveis com o material prospectivo de ensaio de proficiência no que se refere à matriz e concentração, especiação e separação do analito, o resultado para o material de ensaio de proficiência, determinado através de uma função de calibração baseada nos valores certificados dos MRCs, será rastreável aos valores do MRCs e, através deles, a padrões superiores. A incerteza incorporará apenas termos devidos às incertezas dos MRCs e ao erro da repetitividade da análise.

Nota: Esta prática está descrita na ISO 13528 [5]. Ela é idêntica a executar uma medição usando MRCs similares como calibrantes, e poderia, portanto, ser razoavelmente descrita como "medição usando materiais de calibração similares"

Na prática, é difícil determinar se os MRCs são suficientemente similares em todos aspectos ao material de ensaios de proficiência. Se eles não são similares, uma contribuição adicional deve ser incluída no cálculo da incerteza do valor designado. É difícil determinar a magnitude desta contribuição adicional. Como acima, os ensaios de proficiência são mais úteis em setores analíticos onde os materiais de referência não estão disponíveis.

3.2.4 Consenso de laboratórios especialistas

O valor designado é tomado como o consenso de um grupo de laboratórios especialistas que alcançam um entendimento quanto ao material de ensaio de proficiência, através da cuidadosa execução de métodos de referência reconhecidos. Este método é particularmente útil onde parâmetros definidos operacionalmente ("empíricos") são medidos, ou, por exemplo, onde se espera que resultados laboratoriais de rotina sejam consistentes com os resultados de uma população menor de laboratórios, identificados por lei para fins de arbitragem ou regulação. Isto também traz a vantagem de facilitar a verificação cruzada entre laboratórios especialistas, o que ajuda a impedir erros grosseiros.

Na prática, de qualquer maneira, o esforço requerido para chegar a um consenso e a uma pequena incerteza utilizável é mais ou menos o mesmo requerido para certificar um material de referência. Se os laboratórios de referência usassem um procedimento de rotina para analisar o material de ensaio de proficiência, seus resultados tenderiam a não ser melhores na média do que aqueles da maioria dos participantes do ensaio de proficiência propriamente dito. Além disso, como o número de laboratórios de referência disponíveis é intrinsecamente pequeno, a incerteza e/ou variabilidade do consenso de uma subpopulação poderiam ser suficientemente grandes para prejudicar o ensaio de proficiência.

Onde se usa o consenso de laboratórios especialistas, o valor designado e a incerteza associada são avaliados usando-se uma estimativa apropriada da tendência central (geralmente, a média ou uma estimativa robusta). A incerteza do valor designado é então baseada ou nas incertezas relatadas combinadas (caso consistentes), ou na incerteza estatística apropriada combinada com quaisquer termos adicionais requeridos para considerar as incertezas da cadeia de calibração, efeitos matriz, e quaisquer outros efeitos.

3.2.5 Formulação

A formulação consiste na adição de uma quantidade conhecida ou uma concentração conhecida de analito (ou material contendo o analito) a um material base que não o contém. As seguintes circunstâncias devem ser consideradas.

- O material base deve estar efetivamente livre de analito ou sua concentração deve ser conhecida com precisão.
- Pode ser difícil obter homogeneidade suficiente (ver Seção 3.11) quando um analito-traço é adicionado a um material base sólido.
- Mesmo quando a especiação é adequada, o analito adicionado pode se tornar menos ligado à matriz que o analito nativo encontrado em ensaios típicos de materiais, e portanto a recuperação do analito adicionado pode se tornar consideravelmente irreal.

– Caso estes problemas possam ser superados, o valor designado é determinado simplesmente pelas proporções dos materiais usados e as concentrações conhecidas (ou pureza, se um analito puro é adicionado). Sua incerteza é normalmente estimada a partir de incertezas relativas à pureza ou `as concentrações de analito dos materiais usados e incertezas gravimétricas e volumétricas, embora aspectos como conteúdo de misturas e outras mudanças ocorridas durante a mistura também devam ser levadas em consideração, caso significativos. O método é relativamente fácil de executar quando o material do ensaio de proficiência é um líquido homogêneo e o analito está em solução verdadeira. Contudo, pode ser inadequado para materiais naturais sólidos onde o analito já está presente ("nativo" ou "adicionado").

3.2.6 Consenso de participantes

O consenso de participantes é atualmente o método mais amplamente usado para determinação do valor designado: de fato, raramente há uma alternativa, em termos de custo/benefício. A idéia do consenso não é a de que todos os participantes concordem dentro dos limites determinados pela precisão da repetitividade, mas a de que os resultados produzidos pela maioria sejam não-tendenciosos (*unbiased*) e sua dispersão tenha um perfil prontamente identificável. Para se deduzir o valor mais provável do mensurando (isto é, o valor designado) usa-se uma medida apropriada da representatividade (tendência central) dos resultados e (geralmente) usa-se o seu desvio-padrão da média como a estimativa de sua incerteza (ver Seção 3.3).

As vantagens do consenso de participantes incluem o baixo custo, porque o valor designado não requer trabalho analítico adicional. A aceitação de pares é freqüentemente boa entre os participantes porque a nenhum membro ou grupo é conferido um *status* diferenciado. O cálculo do valor é geralmente direto. Por último, uma vasta experiência demonstrou que os valores de consenso geralmente situam-se muito próximos, na prática, de valores de referência confiáveis obtidos através de formulação, consenso de laboratórios especialistas, e valores de referência (sejam de MRCs ou laboratórios de referência).

As principais desvantagens dos valores de consenso de participantes são, em primeiro lugar, que eles não são independentes dos resultados dos participantes e, em segundo lugar, que sua incerteza pode ser muito grande quando o número de laboratórios é muito pequeno. A falta de independência tem dois efeitos em potencial: (i) uma tendência (bias) da população como um todo pode não ser prontamente identificada, na medida em que o valor designado terá o mesmo comportamento da população; (ii) se a maioria dos resultados são tendenciosos (*biased*), os participantes cujos resultados não o sejam podem injustamente receber índices-z (*z-scores*) extremos. Na prática, o primeiro caso é raro, exceto quando se usa o mesmo método em populações pequenas; a existência de várias subpopulações distintas é um problema mais comum. É recomendável que tanto os provedores de ensaios de proficiência quanto os participantes estejam devidamente atentos para estas possibilidades (embora igualmente atentos para a possibilidade de erro em qualquer outro método de designação de valor). A situação geralmente é rapidamente corrigida uma vez identificada. Um dos benefícios dos ensaios de proficiência é que os participantes podem se tornar conscientes de problemas gerais não identificados, assim como daqueles que envolvam laboratórios específicos.

As limitações induzidas por tamanhos pequenos de grupos são freqüentemente mais sérias. Quando o número de participantes é menor do que 15, mesmo a incerteza estatística associada ao consenso (identificada como sendo o desvio-padrão da média) será considerada indesejavelmente alta, e o conteúdo de informação dos índices-z (*z-score*s) será analogamente reduzido.

A despeito das aparentes desvantagens, entretanto, há uma vasta experiência demonstrando que os ensaios de proficiência funcionam muito bem através do uso do consenso, contanto que os organizadores estejam conscientes da possibilidade de eventuais dificuldades e apliquem os métodos adequados de cálculo. Métodos corretos de estimativa do consenso a partir dos resultados dos participantes são devidamente discutidos em detalhe abaixo.

3.3 Estimando o valor designado como o consenso dos resultados dos participantes

3.3.1 Estimativas da tendência central

Se os resultados dos participantes numa rodada são unimodais e, dispersos à parte, aproximadamente simétricos, as várias medidas da representatividade (de tendência central) são praticamente coincidentes. Assim, pode-se com confiança ter uma delas, tal como a moda, a mediana ou a média robusta, como o valor designado.

Precisa-se usar um método de estimação que seja insensível à presença de dispersos e extremidades com muitos dados, a fim de evitar influências indevidas de resultados insatisfatórios, e é por isso que a mediana ou uma média robusta é útil.

A estatística robusta parte da suposição de que os dados são uma amostra de uma distribuição essencialmente de Gauss modificada por extremidades com muitos dados e uma pequena proporção de dispersos. As estatísticas são calculadas atribuindo menor peso aos pontos de dados distantes da média e então compensando por este desconto. Há muitos tipos de estatística robusta [5,11]. A mediana é um tipo simples de média robusta. A média robusta de Huber, obtida através do algoritmo recomendado pelo Comitê de Métodos Analíticos (AMC) [11], e pelas ISO 5725 e ISO 13528 como o "algoritmo A", faz mais uso das informações contidas nos dados do que a mediana o faz, e, consequentemente, na maioria das circunstâncias apresenta um desvio-padrão da média um pouco menor. A mediana, entretanto, é mais robusta quando a freqüência de distribuição é fortemente assimétrica. A média robusta é, portanto, preferida quando a distribuição é aproximadamente simétrica. A moda não é definida com exatidão para amostras de distribuições contínuas, e métodos especiais são necessários para estimá-la. Contudo, a moda pode ser especialmente útil quando resultados bimodais ou multimodais são obtidos (ver Apêndice 3).

Um esquema recomendado para estimar o consenso e sua incerteza é esboçado abaixo. Um elemento de discernimento, baseado no conhecimento de especialistas em química analítica e estatística, é inserido dentro de tal esquema; isto é um ponto forte mais do que um ponto fraco, e é considerado essencial. Isto acontece porque é difícil ou impossível vislumbrar um conjunto de regras que possam ser executadas automaticamente para se obter um consenso adequado a partir de qualquer conjunto de dados arbitrariamente escolhido.

3.3.2 Esquema recomendado para obter o valor de consenso e sua incerteza

O esquema recomendado para se obter através de consenso um valor designado x_a e sua incerteza é definido no procedimento descrito no quadro abaixo. Os fundamentos de certos detalhes são discutidos na Seção 3.3.3. Exemplos de uso deste esquema encontram-se no Apêndice I.

Recomendação 1

a. Excluir dos dados quaisquer resultados que são identificáveis como inválidos (i.é., se expressos nas unidades erradas ou obtidos por uso de um método proscrito) ou que são dispersos extremos, por exemplo., fora da faixa de ±50 % da mediana).

b. Examinar a disposição visual dos resultados restantes, por meio de gráfico de pontos [para conjuntos pequenos de dados (n < 50)], gráficos de barras, ou histograma (para conjuntos maiores). Caso os dispersos tornem indevidamente comprimida a disposição da maioria dos resultados, fazer novo gráfico com os dispersos eliminados. Se a distribuição é, dispersos fora, aparentemente unimodal e aproximadamente simétrica, ir para (c), senão ir para (d).

c. Calcular a média robusta $\hat{m}_{rob}$ e o desvio-padrão $\hat{s}_{rob}$ dos n resultados. Se $\hat{s}_{rob}$ é menor que 1,2 s_p, então usar $\hat{m}_{rob}$ como o valor designado x_a e $\hat{s}_{rob}/\sqrt{n}$ como sua incerteza-padrão. Se $\hat{s}_{rob} > 1{,}2\ s_p$, ir para (d).

d. Fazer uma estimativa do estimador de intensidade da distribuição dos resultados usando intensidades (*kernel*s) normais com uma amplitude h de $0{,}75\ s_p$. Se isto resultar em um estimador de intensidade unimodal e aproximadamente simétrico, e a moda e mediana forem quase coincidentes, então usar $\hat{m}_{rob}$ como o valor designado x_a e $\hat{s}_{rob}/\sqrt{n}$ como sua incerteza-padrão. Senão, ir para (e).

e. Se as modas secundárias puderem ser atribuídas com segurança a resultados dispersos, e se contribuem com menos de 5 % da área total, então usar $\hat{m}_{rob}$ como o valor designado x_a e $\hat{s}_{rob}/\sqrt{n}$ como sua incerteza-padrão. Senão, ir para (f).

f. Se as modas secundárias contribuem consideravelmente para com a área da intensidade, considerar a possibilidade de que duas ou mais populações discrepantes estejam representadas nos resultados dos participantes. Caso seja possível inferir a partir de informações independentes (ex: detalhes dos métodos analíticos dos participantes) que uma das modas é correta e as outras incorretas, usar a moda selecionada como o valor designado x_a e seu desvio-padrão da média como a sua incerteza-padrão. Senão, ir para (g).

g. Se falharem os métodos acima, abandonar as tentativas de determinar um valor de consenso e não relatar índices *(scores)* individuais de desempenho de laboratórios para a rodada. Poderá ainda ser útil, entretanto, fornecer aos participantes um sumário estatístico do conjunto de dados como um todo.

3.3.3 Notas sobre os fundamentos do esquema de determinação do valor designado

Os fundamentos do esquema acima são os que seguem:

O uso de $\hat{\sigma}_{rob}/\sqrt{n}$ como a incerteza-padrão do valor designado está sujeita a objeção em bases teóricas, porque a influência de alguns dos n resultados é parcialmente descontada ao se calcular $\hat{\sigma}_{rob}$ e sua distribuição amostral é complexa. Ele é, entretanto, um dos métodos recomendados na ISO 13528. Na prática, $u(x_a) = \hat{\sigma}_{rob}/\sqrt{n}$ é apenas usado como uma indicação grosseira da adequação do valor designado, e a objeção teórica é de pouco interesse.

Em (b) acima, espera-se $\hat{\sigma}_{rob} \approx \sigma_p$, já que os participantes estarão tentando obter adequação aos propósitos. Se detectarmos que $\hat{\sigma}_{rob} > 1{,}2\,\sigma_p$, é razoável supor que ou os laboratórios estão tendo dificuldades para obter a precisão de reprodutibilidade requerida nos resultados de uma população, ou que uma ou mais populações dispersas podem estar representadas nos resultados. Uma densidade de *kernel* pode ajudar a decidir entre estas duas possibilidades. Se o último caso resulta ou não em duas (ou mais) modas depende da separação das médias das populações e do número de resultados em cada amostra.

Usar uma amplitude h de $0{,}75\,\sigma_p$ para interpretar densidades de *kernel* é um meio-termo que inibe a incidência de modas artificiais sem aumentar indevidamente a variância da densidade de *kernel* em relação a $\hat{\sigma}_{rob}$.

3.4 Incerteza do valor designado

Se existe uma incerteza u(x_a) no valor designado x_a, e um participante cujo desempenho está de acordo com o desvio-padrão σ_p que define adequação aos propósitos, a incerteza no desvio de x_a pelo participante seria $\sqrt{u^2(x_a) + \sigma_p^2}$, de maneira que se poderia esperar a incidência de índices-z (*z-scores*) com uma distribuição outra do que N(0, 1). É, portanto, adequado comparar $u^2(x_a)$ com σ_p^2 para se certificar que a primeira não está tendo um efeito adverso sobre os índices-z (*z-scores*). Por exemplo, se $u^2(x_a) = \sigma_p^2$, os índices-z (*z-scores*) seriam ampliados cerca de 1,4 vezes, o que seria um resultado inaceitável. Por outro lado, se $u^2(x_a) > \sigma_p^2$, o fator de ampliação seria da ordem de 1,05, cujo efeito seria insignificante para fins práticos. Assim, é recomendado que os índices-z (*z-scores*) não sejam apresentados aos participantes sem observações caso se descubra que $u^2(x_a) > 0{,}1\ \sigma_p^2$. (O fator 0,1 é de magnitude adequada, mas seu valor exato é essencialmente arbitrário e convém que seja considerado pelo provedor do programa). Se a desigualdade for apenas ligeiramente ultrapassada (mas não muito), o programa poderia divulgar os índices-z (*z-scores*) s com uma observação de advertência anexada a eles, por exemplo, rotulando-os como (provisórios), com uma explicação conveniente. Portanto, os provedores de ensaios de proficiência precisariam indicar um valor adequado de l na expressão $u^2(x_a) = l\ \sigma_p^2$, acima do qual nenhum índice-z (*z-score*) seria calculado.

A norma ISO 13528 [5] refere-se a um índice " z' " modificado que seria dado por $z' = \dfrac{x - x_a}{\sqrt{u^2(x_a) + \sigma_p^2}}$ que poderia ser usado quando a incerteza do valor designado for significativo. Entretanto, z' *não é* recomendado para uso neste protocolo. Embora ele tendesse a fornecer valores similares a índices-z (*z-scores*) em dispersão, o uso de z' iria mascarar o fato de que a incerteza do valor designado é inadequadamente alta. Portanto, a recomendação atual é a seguinte:

Recomendação 2
Convém que o provedor de ensaios de proficiência indique um multiplicador $0,1 < l < 0,5$ adequado para o programa e, tendo estimado $u^2(x_a) + s_p^2$ para uma rodada, aja como segue:

- se $u^2(x_a) + s_p^2 \leq 0,1$, divulgar índices-z (*z-scores*) sem observações;
- se $0,1 < u^2(x_a) + s_p^2 \leq l$, divulgar índices-z (*z-scores*) com observações (tais como "índices-z (*z-scores*) provisórios");
- se $u^2(x_a) + s_p^2 > l$, não divulgar índices-z (*z-scores*).

Nota: Na desigualdade $0,1 < l < 0,5$, os limites podem ser ligeiramente modificados a fim de atender a requisitos exatos de programas específicos.

3.5 Determinação do desvio-padrão para avaliação da proficiência

O desvio-padrão para avaliação da proficiência. s_p é um parâmetro que é usado para prover um escalonamento para os desvios $(x - x_a)$ do valor designado por parte do laboratório e assim definir um índice-z (*z-score*) (Seção 3.1). Há várias maneiras pelas quais o valor de um parâmetro pode ser determinado, e seus méritos relativos são abaixo discutidos.

3.5.1 Valor determinado por adequação aos propósitos

Neste método, o provedor do programa de proficiência determina um nível de incerteza que é amplamente aceito como apropriado pelos participantes e usuários finais de dados do setor de aplicação dos resultados, e o define em termos de s_p . Por "apropriado", entende-se que a incerteza é pequena o suficiente para que as decisões baseadas nos dados raramente sejam incorretas, mas não tão pequena que os custos de análise sejam indevidamente altos. Uma definição sugerida desta “adequação aos propósitos” é a de que ela deve abranger a incerteza que minimiza os custos combinados de análise e as conseqüências financeiras associadas a decisões incorretas multiplicadas pela sua probabilidade de ocorrência [10]. Deve-se enfatizar que s_p não representa aqui uma estimativa de como está o desempenho dos laboratórios, mas sim como deveria estar de maneira a poderem cumprir seu compromissos com seus clientes. Convém que o valor numérico do parâmetro seja tal que os índices-z (*z-scores*) resultantes possam ser interpretados com referência à distribuição normal padrão. Ele será provavelmente determinado pelo discernimento profissional exercido pelo comitê assessor do programa. Em alguns setores analíticos, já existe um padrão reconhecido em uso para “adequação aos propósitos”. Por exemplo, no setor de alimentos, a função de *Horwitz* é freqüentemente considerada uma definição de “adequação aos propósitos”, tanto quanto é considerada simplesmente descritiva [13].

Seja como for que se chegue ao parâmetro, este terá que ser definido e divulgado antes da distribuição dos materiais de ensaio de proficiência, de forma que os participantes possam verificar se os seus procedimentos analíticos são adequados a ele. Em alguns programas, a gama possível de concentrações de analito é pequena e um nível de incerteza individual pode ser especificado para cobrir todas as eventualidades. Os problemas surgem em situações onde a concentração de analito pode variar dentro de uma ampla gama. Como o valor designado não é conhecido antecipadamente pelos participantes, o critério de “adequação aos propósitos” tem que ser especificado como uma função da concentração. As abordagens mais comuns são:

- Especificar o critério como um desvio-padrão relativo (DPR). Valores específicos de s_p são então obtidos multiplicando-se esse DPR pelo valor designado.
- Quando há um limite inferior de interesse em um resultado analítico, estabelecer um DPR aplicável à faixa especificada em conjunto com um valor limite (inferior) para s_p . Por exemplo, na determinação da concentração de chumbo em vinho, seria prudente focar um DPR de 20% numa ampla faixa de concentrações de analito, mas em concentrações muito abaixo da concentração máxima permitida $x_{máx}$ tal nível de precisão não seria nem necessário nem econômico. Esse fato poderia ser reconhecido

expressando-se o critério de "adequação aos propósitos"

$$\sigma_p = x_{\max} / m + 0{,}2x_a \quad (2)$$

onde *f* é uma constante adequada. Se *f* fosse considerado como 4, por exemplo, σ_p nunca seria menor que $x_{\max}/4$.

Especificar uma expressão geral da "adequação aos propósitos", tal como a função de *Horwitz* [13], a saber, (usando notação atual):

$$\sigma_p = 0{,}02\,x_a^{0{,}8495} \quad (3)$$

onde x_a e σ_p são expressos em fração mássica. Notar que a relação de *Horwitz* original perde aplicabilidade em concentrações aproximadamente menores que 10 ppb (ppb = 10^9 fração mássica), e uma forma modificada da função tem sido recomendada [14].

3.5.2 Valor legalmente definido

Em alguns casos, um desvio-padrão máximo da reprodutibilidade dos resultados analíticos para um fim específico é estabelecido por lei ou acordo internacional. Este valor pode ser usado como um valor para σ_p. Analogamente, se foi estabelecido um limite de erro permitido, este pode ser usado para se estabelecer σ_p, através, por exemplo, da divisão pelo adequado valor *t* de Student, caso um nível de confiança também seja disponível. Entretanto, pode ser ainda preferível usar um valor de σ_p menor que o limite legal. Isto é uma questão a ser decidida pelo provedor e o comitê assessor do programa de ensaios de proficiência.

3.5.3 Outras abordagens

Em alguns programas de ensaios de proficiência, o índice *(score)* não é baseado na "adequação aos propósitos", o que diminui enormemente o seu valor de uso. Embora tais métodos de índice *(score)* sejam abordados pela norma ISO Guia 43 [4], (e discutidos na versão anterior deste Protocolo Harmonizado [1]), eles não são aqui recomendados para ensaios de proficiência química. Especificamente, existem duas versões de tais sistemas de índice *(score)*. Numa delas, o valor de σ_p é determinado pelo discernimento de especialistas quanto ao desempenho do laboratório para o tipo de análise que está sendo considerada. De fato, o desempenho dos laboratórios pode ser melhor ou pior que a "adequação aos propósitos", de forma que o sistema de índice *(score)* nesse caso apenas nos informa quais laboratórios estão divergentes dos outros participantes, mas não nos informa se alguns deles são bons o suficiente. Outra versão deste método, aparentemente mais reconhecida porque baseada em conceitos estatísticos padrão, é usar σ_p como desvio-padrão robusto dos resultados dos participantes de uma determinada rodada. O resultado de tal estratégia é que em cada caso, aproximadamente 95% dos participantes recebem um índice-z (*z-score*) aparentemente aceitável. Isto é um resultado reconfortante tanto para os participantes quanto para os provedores de programas, mas, de novo, serve apenas para identificar resultados divergentes dos outros. Há como dificuldade adicional pelo fato de que o valor usado para σ_p variará de rodada para rodada de forma que não há uma base de comparação estável de índice (*score*) entre as rodadas. Embora o método possa ser melhorado pelo uso de um valor fixo derivado da combinação dos resultados de várias rodadas, ele ainda não faz nenhum incentivo aos laboratórios que produzem resultados "inadequados para os propósitos" a melhorar seu desempenho.

Pode haver circunstâncias onde é justificável que um programa de ensaios de proficiência não forneça orientação sobre a adequação aos propósitos. Este é o caso quando os participantes conduzem a sua rotina diária sobre uma variedade de propósitos diferentes, de forma que não pode existir uma "adequação aos propósitos" aplicável globalmente. Em tais condições, seria melhor para o provedor de ensaios de proficiência não fornecer nenhum índice *(score)* mas apenas fornecer um valor designado (com sua incerteza) e, talvez, o erro do laboratório. (Este último é às vezes fornecido em termos de erro relativo, o assim

chamado "índice Q" . É recomendável que qualquer índice *(score)* fornecido seja claramente identificado como "apenas para uso informal", a fim de minimizar a incidência de julgamentos incorretos com base nos índices *(scores)*, da parte de avaliadores de acreditação ou clientes em potencial. Participantes individuais de tais programas têm então que providenciar os seus próprios critérios de "adequação aos propósitos", e um roteiro para implementar isto é fornecido abaixo (ver Seção 3.6 e Apêndice 6).

Recomendação 3

Convém que, sempre que possível, o programa de ensaios de proficiência use para σ_p o desvio-padrão para avaliação da proficiência, um valor que reflita a adequação aos propósitos para o setor. Se não existe um nível único que seja adequado de forma geral, convém que o provedor se abstenha de calcular índices *(scores)*, ou mostre claramente nos relatórios que os índices *(scores)* são para uso descritivo informal apenas e não para serem considerados um índice de desempenho dos participantes.

3.5.4 Índice-z (z-score) modificado para requisitos individuais

Em alguns programas de ensaios de proficiência, o índice *(score)* não é baseado na "adequação aos propósitos". O provedor do programa calcula um índice *(score)* a partir dos resultados dos participantes apenas (isto é, sem referência externa a requisitos reais). Alternativamente, um participante pode achar que o critério de adequação aos propósitos usado pelo provedor do programa é inadequado para certas classes de trabalho desenvolvidas pelo laboratório. De fato, não seria incomum para um laboratório ter alguns clientes que desejam a determinação do mesmo analito no mesmo material, mas cada um deles tendo um valor de incerteza diferente.

Em programas de ensaios de proficiência que operam numa destas duas bases, os participantes podem calcular índices *(scores)* com base nos próprios requisitos de adequação aos propósitos. Isto pode ser obtido de maneira direta. É recomendável que o participante acorde com o cliente um critério específico de adequação aos propósitos σ_{ffp} para cada aplicação específica, usando-o para calcular o índice-z (*z-score*) modificado correspondente, dado por:

$$z_L = \frac{(x - x_a)}{\sigma_{ffp}} \tag{4}$$

a fim de substituir o índice-z (*z-score*) convencional [15]. O critério σ_{ffp} pode ser expresso como uma função da concentração, caso necessário. Convém que seja usado como o valor sigma num índice-z (*z-score*), isto é, na forma de uma incerteza-padrão que representa a adequação aos propósitos que foi acordada. Se houver vários clientes com diferentes requisitos de exatidão, poderá haver vários índices *(scores)* válidos derivados de qualquer resultado individual. Os índices-z (*z-scores*) modificados podem ser interpretados exatamente da maneira recomendada para índices-z (*z-scores*) (ver Apêndice 6).

3.6 Dados de participante relatados com incerteza

Este Protocolo não recomenda o relato da incerteza de medição dos participantes junto com o os resultados. Esta recomendação é consistente com o ISO Guia 43. De fato, relativamente poucos programas de ensaios de proficiência para química analítica atualmente requerem que os resultados dos participantes sejam acompanhados da estimativa de incerteza. Isto acontece principalmente porque se assume, depois de cuidadosa ponderação de especialistas, que os programas normalmente estabelecem um valor σ_p que representa a adequação aos propósitos de todo um setor de aplicação. Este requisito de incerteza ótima é, portanto, implícito ao programa. Espera-se que os participantes tenham um desempenho consistente com essa especificação e, portanto (neste contexto), não precisem relatar as incertezas de maneira explícita. Aqueles cujo desempenho estiver conforme ao requerimento do programa geralmente receberão índices-z (*z-scores*) na faixa de ±2. Aqueles participantes que tiverem incertezas significativamente subestimadas muito mais provavelmente receberão índices-z (*z-scores*) qualificados como "inaceitável". Em outras palavras, espera-se que as incertezas corretamente estimadas sejam na maioria similares ao valor σ_p e que as subestimadas tenderão a resultar em índices-z (*z-scores*) *insatisfatórios*. Em tais circunstâncias, o relato da incerteza nada adiciona ao valor do programa. necessários para a melhora do desempenho analítico de rotina.

Entretanto, as circunstâncias esboçadas acima podem não ser aplicáveis universalmente. É recomendável, portanto, que laboratórios que buscam os seus próprios critérios de adequação sejam julgados por critérios individuais ao invés do uso genérico do valor σ_p para o programa. Ademais, dados sobre incerteza são cada vez mais requeridos por clientes dos laboratórios, e convém que os laboratórios estejam verificando seus procedimentos para tanto. As seguintes seções discutem três aspectos importantes relativos ao uso das incertezas do participante: a determinação dos valores de consenso, o uso dos índices *(scores)* como uma verificação da incerteza relatada e o uso de incerteza do participante ao avaliar a "adequação individual aos propósitos".

3.6.1 Valores de consenso

Quando estimativas de incerteza estão disponíveis, provedores de programas talvez precisem ponderar sobre a melhor maneira de identificar o consenso quando os participantes relatam dados sobre incertezas e como o consenso dessa incerteza é melhor estimado. A versão ideal do problema é estabelecer um consenso e suas incertezas a partir de um conjunto de estimativas não tendenciosas (*unbiased estimates*) de um mensurando, cada uma com uma diferente incerteza. A verdade é que: (a) freqüentemente há resultados discordantes entre aqueles relatados (isto é, os dados abrangem amostras de distribuições com médias diferentes); e (b) as estimativas de incerteza são freqüentemente incorretas e, em particular, aquelas agregadas a resultados dispersos tendem a ser demasiadamente pequenas.

Atualmente, não há métodos solidamente estabelecidos para prover estimativas robustas de médias ou dispersão para dados interlaboratoriais com incertezas variáveis. Esta área está, entretanto, em franco desenvolvimento, e várias propostas interessantes têm sido discutidas [16,17]. Por exemplo, os métodos baseados na estimativa de densidade de *kernel* [12] atualmente aparentam ser produtivos.

Felizmente, a maioria dos participantes de um dado programa busca requisitos similares e espera-se que forneçam incertezas também similares. Nestas circunstâncias, as estimativas ponderadas e não ponderadas da tendência central são muito similares. Estimativas robustas não ponderadas devem, portanto, ser aplicadas com a vantagem de menor sensibilidade a subavaliações substanciais da incerteza ou a dispersos distantes.

Dado que tais tópicos ainda estão em desenvolvimento e que as estimativas de incerteza são ligeiramente não-confiáveis, recomenda-se que métodos robustos não ponderados sejam usados para o cálculo do valor designado.

Recomendação 4

Mesmo quando as estimativas de incerteza estão disponíveis, convém usar métodos robustos não ponderados (isto é., métodos que não levam em consideração as incertezas individuais) para se obter o valor de consenso e suas incertezas, de acordo com os métodos descritos nas seções 3.3 e 3.4.

3.6.2 O índice zeta

A ISO 13528 define o índice *zeta* (ζ) para se atribuir índices *(scores)* a resultados e incertezas relatadas, conforme segue:

$$z = \left(x - x_a\right) \Big/ \sqrt{u^2(x) + u^2(x_a)} \tag{5}$$

onde $u(x)$ é a incerteza-padrão relatada no valor reportado x e $u(x_a)$ a incerteza-padrão do valor designado. O índice *zeta* fornece a indicação de que a estimativa de incerteza do participante é consistente com o desvio observado de um valor designado. A interpretação é similar à interpretação dos índices-z (*z-scores*); é recomendável que valores absolutos acima de 3 sejam considerados como motivo para investigação mais aprofundada . O motivo poderia ser a subavaliação da incerteza $u(x)$, mas poderia também ser um erro grosseiro fazendo com que o desvio $x - x_a$ seja muito grande. É de se esperar que o último caso acima resulte num alto índice-z (*z-score*), de forma que é importante considerar os índices z e *zeta* em conjunto. Observe também que índices *zeta* persistentemente baixos por um período de tempo poderiam indicar super avaliação da incerteza.

Nota: A ISO 13528 define métodos de índice *(score)* adicionais que utilizam a incerteza expandida;

referência à ISO 13528 é recomendada caso isto seja considerado adequado pelo comitê assessor do programa.

3.6.3 Atribuindo índices (scores) a resultados relatados com incerteza

É fácil para um participante usar o índice *zeta* para verificar suas próprias estimativas de incerteza. No momento, entretanto, estamos considerando medidas tomadas pelos organizadores de programas de ensaios de proficiência.

A questão em pauta é se convém que o provedor do programa (ao invés de participantes individuais) tente levar a incerteza em consideração na conversão dos resultados brutos em índices *(scores)*. Não há nenhuma dificuldade específica em fazer isso—trata-se meramente de observar se o resultado final é útil ou não. A avaliação dos benefícios depõe contra os programas que calculam os índices *zeta*. Todos os programas são incentivados a fornecer um diagrama mostrando os resultados e índices *(scores)* dos participantes. Tal diagrama, baseado em índices *zeta*, seria ambíguo porque os resultados não poderiam ser representados de maneira útil num gráfico bi-dimensional. Um índice *zeta* específico (digamos, -3,7) poderia se originar tanto de um erro grande e uma grande incerteza, quanto de um pequeno erro e uma incerteza proporcionalmente pequena. Além disso, o organizador do programa não tem meios para julgar se o valor de incerteza fornecido por um participante é adequado às suas necessidades, de forma que os índices *zeta* assim produzidos teriam uma utilidade desconhecida na avaliação dos resultados dos participantes.

Recomendação 5

É recomendável que programas não forneçam índices *zeta* a não ser que haja razões especiais para isso. Quando um participante apresenta requisitos inconsistentes com aqueles do programa, o participante pode calcular índices *zeta* ou equivalentes.

3.7 Atribuindo índices *(scores)* a resultados próximos do limite de detecção

Muitas tarefas analíticas envolvem a medição de concentrações de analito que estão próximas ao limite de detecção do método, ou mesmo que estão exatamente em zero. Convém que o ensaio de tais métodos quanto à proficiência espelhe a vida real; convém que os materiais de ensaio contenham concentrações de analito tipicamente baixas. Contudo, há dificuldades em aplicar o método usual de índices-z (*z-scores*) aos resultados de tais ensaios. Estas dificuldades são parcialmente causadas pelas seguintes práticas de registro de dados:

- Muitos analistas ao obterem um resultado de valor baixo, registrarão um resultado como "não detectado" ou "menos que c_L ", onde c_L é um limite arbitrariamente determinado. Tais resultados, embora possivelmente adequados aos propósitos, não podem ser convertidos num índice-z (*z-score*). Índices-z (*z-scores*) requerem que o resultado analítico se situe numa escala de intervalo ou numa escala de razão. Substituir o resultado "assumido" de forma arbitrária (por exemplo, por zero ou a metade do limite de detecção) não é recomendado.
- Alguns programas de ensaios de proficiência evitam tal dificuldade simplesmente não processando tais resultados como "não detectado". Caso muitos participantes estejam trabalhando próximo aos seus limites de detecção, independentemente de fornecerem um resultado "não detectado", torna-se difícil estimar um consenso válido para o valor designado. A distribuição dos resultados aparentemente válidos tende a apresentar uma forte assimetria positiva, e a maioria dos tipos de valores médios tende a apresentar uma tendência (*bias*) pronunciada.

Estas dificuldades podem ser contornadas trabalhando-se com concentrações ligeiramente maiores do que aquelas tipicamente encontradas nos materiais de interesse. Esta prática não é inteiramente satisfatória, porque as amostras são então irreais. Se os participantes registraram o valor real obtido, mais a (correta) incerteza do resultado, seria possível em princípio estimar um consenso válido com uma incerteza. Embora isto seja recomendado onde for possível, outros fatores tais como práticas estabelecidas nos requisitos de relatórios de clientes a tornam uma prática improvável nas análises de rotina. Portanto, parece que os índices-z (*z-scores*) podem ser usados em baixas concentrações apenas quando todas as condições abaixo estão aplicadas:

- Os participantes registram os reais resultados obtidos.
- O valor designado é independente dos resultados. Isto poderia ser possível caso se soubesse que o valor designado é zero ou muito baixo, ou então caso ele possa ser determinado por formulação ou por um laboratório de referência (ver seção 3.2).

• O desvio-padrão para avaliação da proficiência é um critério de adequação aos propósitos independente;
seu valor poderia então ser predeterminado, isto é, independente dos resultados dos participantes. Isto seria relativamente simples de fazer acontecer.

No presente, não existem sistemas de índices *(scores)* bem estabelecidos para resultados de valor baixo em ensaios de proficiência, e o assunto ainda está em discussão. Caso os resultados sejam essencialmente binomiais ($\leq x$ ou $>x$), então poderia se conceber um sistema de índice *(score)* baseado na proporção de resultados corretos, mas ele tenderia a ser menos rico em informações do que o sistema de índices-z (*z-scores*). Um sistema misto, (capaz de lidar com uma mistura de resultados binomiais, ordinais e quantitativos) ainda não disponível.

3.8 Cautela no uso de índices-z (*z-scores*)

Os usos adequados dos índices-z (*z-scores*) pelos participantes e usuários finais são discutidos em detalhe nos Apêndices 6 e 7. As seguintes palavras de cautela são dirigidas aos provedores.

É comum que diferentes análises sejam requeridas em cada rodada de um ensaio de proficiência. Embora cada ensaio individual forneça informação útil, é tentador determinar uma única nota de mérito que resumirá o desempenho geral do laboratório em determinada rodada. Existe o perigo de que tal índice *(score)* combinado seja mal interpretado ou usado indevidamente por não especialistas, particularmente fora do contexto de índices *(scores)* individuais. Portanto, não se recomenda a determinação geral de usar índices *(scores)* combinados em relatórios para participantes, mas reconhece-se que tais índices *(scores)* podem ter aplicações específicas, caso sejam baseados em sólidos princípios estatísticos e divulgados com uma nota adequada de advertência. Os procedimentos que podem ser usados são descritos no Apêndice 4.

Enfatiza-se em particular que existem limitações e fraquezas em qualquer programa que utilize índices (*scores*) combinados de análises dissimilares. Se um índice *(score)* específico entre os vários produzidos por um laboratório for disperso, o índice *(score)* combinado pode muito bem não ser disperso. Em alguns aspectos, isto é um atributo útil, no sentido de que um lapso em uma determinada análise tem um peso menor no índice *(score)* combinado. Entretanto, existe o perigo de que o laboratório seja consistentemente falho apenas em uma análise em particular, e freqüentemente relate um valor inaceitável para aquela análise em sucessivas rodadas do ensaio. Este fator pode muito bem ser mascarado pela combinação de índices *(scores)*.

3.9 Classificação, ordenação e outras avaliações dos dados de proficiência

A classificação de laboratórios não é o foco do ensaio de proficiência, devendo ser evitada pelos provedores, pois causa mais confusão do que esclarecimento. A substituição de uma medição contínua tal como o índice z (*z-score*) por algumas poucas classes identificadas não tem recomendação do ponto de vista cientifico: a informação não é aproveitada. Consequentemente, a classificação não é recomendada em ensaios de proficiência. Limites de decisão baseados em índices z (*z-scores*) podem ser usados como orientação onde necessário. Por exemplo, um índice z (*z-score*) fora da faixa de ±3 poderia ser visto como indicando a necessidade de uma investigação que levasse, quando necessário, a modificações nos procedimentos. Mesmo assim, tais limites são arbitrários. Convém que um índice *(score)* de 2,9 sempre preocupe tanto quanto um índice *(score)* de 3,1. Além disso, estes são aspectos mais para participantes individuais do que para provedores de programas.

Ordenar os laboratórios com base nos seus índices z (*z-scores*) absolutos obtidos numa rodada de um programa, para organizar uma “lista de iguais”, é ainda mais inadequado do que a classificação. A posição relativa de um participante é ainda mais variável de rodada para rodada do que os índices *(scores)* dos quais ela deriva, e é muito improvável que o laboratório com o menor índice *(score)* absoluto numa rodada seja realmente “o melhor".

Recomendação 6

Convém que os provedores de programas de proficiência, participantes e usuários finais evitem a classificação e ordenação de laboratórios com base em seus índices z (*z-scores*).

3.10 Freqüência das rodadas

A freqüência de distribuição adequada é o resultado de uma série de fatores entre os quais os mais importantes são:

- a dificuldade de executar um CQ analítico efetivo;
- a capacidade do laboratório de processar as amostras de ensaio;
- a consistência dos resultados no campo específico de trabalho abrangido pelo programa;
- a relação custo/benefício do programa;
- a disponibilidade de MRCs no setor analítico específico; e
- o ritmo de mudança dos requisitos analíticos, metodologia, instrumentação, e pessoal no setor de interesse.

As evidências da influência da freqüência das rodadas na eficácia dos ensaios de proficiência são muito esparsas. Apenas um estudo confiável sobre a freqüência foi relatado [18], e que demonstrou (num programa específico) que alterar a freqüência das rodadas de 3 para 6 ao ano não teve efeito significativo (benéfico ou o contrário) sobre o desempenho dos participantes.

Na prática, a freqüência provavelmente será entre uma vez a cada duas semanas a uma vez a cada quatro meses. Uma freqüência maior do que uma vez a cada duas semanas poderia causar problemas relacionados ao tempo de giro das amostras e dos resultados. Isto poderia também incentivar a crença de que o programa de ensaios de proficiência pode ser usado como um substituto do Controle da Qualidade Interno (CQI), uma idéia que deve ser fortemente desencorajada. Se o período entre as distribuições se estender muito além de 4 meses, haverá atrasos inaceitáveis na identificação e correção dos problemas analíticos, e o impacto do programa sobre os participantes poderá ser pequeno. Há pouco valor prático, no que concerne o trabalho analítico de rotina, em ensaios de proficiência implementados em freqüência menor do que duas vezes ao ano.

3.11 Ensaios de homogeneidade e estabilidade suficientes

3.11.1 Ensaios de "homogeneidade suficiente"

Os materiais preparados para ensaios de proficiência e outros estudos interlaboratoriais são geralmente heterogêneos em algum grau, apesar dos melhores esforços para assegurar homogeneidade. Quando um lote de tal material é dividido para distribuição entre vários laboratórios, as unidades produzidas variam ligeiramente entre si quanto à composição. Este protocolo requer que esta variação seja suficientemente pequena para o propósito. Um procedimento recomendado é descrito no Apêndice 1. Os fundamentos deste procedimento são discutidos nos próximos parágrafos.

Quando se efetuam ensaios para a assim chamada "homogeneidade suficiente" em tais materiais, está-se buscando demonstrar que esta variação entre a composição das unidades distribuídas (caracterizada pelo desvio-padrão amostral σ_{sam}) é insignificante em comparação à variação introduzida pelas medições conduzidas pelos participantes no ensaio de proficiência. Como se espera que o desvio-padrão da variação interlaboratorial em ensaios de proficiência seja estimada por σ_p , o "desvio-padrão para avaliação da proficiência", é natural que se use o critério como um valor de referência. O Protocolo Harmonizado de 1993 [1] requeria que o desvio-padrão amostral estimado s_{sam} fosse menor que 30% do desvio-padrão alvo σ_p, isto é, $s_{sam} < \sigma_{all}$, onde o desvio-padrão amostral permitido $\sigma_{all} = 0{,}3\,\sigma_p$.

Esta condição, quando atendida, era chamada "homogeneidade suficiente". Naquele limite, o desvio padrão dos escores de índices-z resultantes seria inflacionado pela heterogeneidade em algo ligeiramente menor que 5% (por exemplo, de 2,0 para 2,1), o que era considerado aceitável. Se a condição não fosse atendida, os escores de índices-z refletiriam, num nível inaceitável, a variação no material assim como a variação no desempenho do laboratório. Os participantes dos programas de ensaios de proficiência precisam se assegurar de que as unidades distribuídas de material de ensaio são suficientemente similares, e este

requisito geralmente demanda ensaios.

O ensaio especificado demandava a seleção aleatória de dez ou mais unidades depois que o material supostamente homogêneo tinha sido dividido e embalado em amostras separadas para distribuição. O material de cada amostra era então analisado em duplicata, sob condições de repetitividade aleatórias (isto é, todas em uma corrida) usando um método com suficiente precisão analítica. O valor de σ_{sam} é então estimado a partir dos quadrados das médias depois de análise da variância de uma via (ANOVA).

Ensaios de homogeneidade suficiente nunca são inteiramente satisfatórios na prática. O principal problema é que, por causa dos altos custos da análise, o número de amostras usadas para ensaios será pequeno. Isto torna a contribuição do ensaio estatístico (isto é, a verificação da probabilidade de rejeição do material quando ele é de fato heterogêneo), relativamente pequena. Um problema adicional é que a heterogeneidade é inerentemente não uniforme, e unidades de distribuição discrepantes podem estar sub-representadas entre aquelas selecionadas para o ensaio. Os ensaios de homogeneidade devem ser considerados essenciais, mas não garantidos ou imunes a erros.

Entretanto, uma vez que a homogeneidade suficiente é uma suposição prévia razoável, (porque os provedores de programas de ensaios de proficiência fazem de tudo para assegurá-la), e uma vez que o custo de fazer ensaios para verificá-la é alto, é sensato enfatizar como principal medida que se procure evitar "Erros Tipo 1" (isto é, falsa rejeição de material satisfatório). Tal medida teria o efeito tornar um ensaio modificado menos inclinado a rejeitar boas amostras.

Para fazer um ensaio de homogeneidade suficiente, têm-se que estimar σ_{sam} a partir dos resultados de um experimento replicado aleatoriamente através do uso de ANOVA. No experimento, cada unidade de distribuição selecionada é separadamente homogeneizada e analisada em duplicata. Muito depende da qualidade dos resultados analíticos. Se o método analítico é suficientemente preciso, σ_{sam} pode ser estimado com confiabilidade, e qualquer falta de homogeneidade suficiente pode ser detectada com probabilidade razoavelmente alta. De fato, o ensaio pode até ser muito sensível. O material pode ser significativamente heterogêneo estatisticamente, mas a variância amostral pode ser insignificante em relação a σ_p . Entretanto, se o desvio padrão analítico σ_{an} não é pequeno, importantes variações nas amostras podem ser mascaradas pela variação analítica. Pode-se obter um resultado não significativo quando se ensaia para heterogeneidade, não porque não esteja presente, mas porque o ensaio não consegue detectá-la.

O Protocolo Harmonizado de 1993, embora especificasse a necessidade de uma precisão analítica suficientemente boa, não especificava nenhum limite numérico para σ_{an}, mas a discussão acima indica que é desejável fazê-lo. Ao se estabelecer este valor, tem que haver uma troca entre os custos de especificar métodos analíticos muito precisos e o risco de falhar em detectar variações amostrais importantes. Recomenda-se assim que a precisão da repetitividade analítica do método usado no ensaio de homogeneidade satisfaça

$$\sigma_{an} / \sigma_p < 0{,}5$$

Entretanto, deve-se reconhecer que eventualmente pode ser impraticável cumprir este requisito, e assim é necessário um procedimento estatístico que forneça um resultado significativo independentemente do valor de σ_{an}.

Recomendação 7
Convém que a precisão (da repetitividade) analítica do método usado no ensaio de homogeneidade satisfaça $\sigma_{an} / \sigma_p < 0{,}5$ onde σ_{an} é o desvio padrão da repetitividade adequado ao ensaio de homogeneidade.

3.11.2 O novo procedimento estatístico

Em vez de expressar o critério para homogeneidade suficiente em termos da variância amostral estimada s^2_{sam}, como no Protocolo Harmonizado de 1993, é mais lógico impor um limite à variância amostral verdadeira σ^2_{sam} [19]. É esta variância amostral verdadeira que é mais pertinente à variabilidade nas amostras (não ensaiadas) enviadas aos laboratórios. Assim , o novo critério para homogeneidade suficiente é que a variância amostral σ^2_{sam} não deve exceder uma grandeza permitida $\sigma^2_{all} = 0{,}09\ \sigma^2_p$ (isto é, σ_{all} =

0,3 σ_p). Então, nos ensaios de homogeneidade faz sentido testar a hipótese $\sigma_{sam}^2 \leq \sigma_{all}^2$ contra a alternativa $\sigma_{sam}^2 > \sigma_{all}^2$. (O teste-*F* costumeiro numa ANOVA de uma via teste a hipótese mais estrita $\sigma_{sam}^2 = 0$ contra a alternativa $\sigma_{sam}^2 > 0$, que poderia prover evidência de que há uma variação amostral, mas não necessariamente que ela é inaceitavelmente grande). O novo procedimento é concebido para acomodar este requisito e as outras dificuldades mencionadas acima. O procedimento completo e seu exemplo de aplicação são descritos no Apêndice 1.

Recomendação 8

Empregar um teste explícito da hipótese H: $\sigma_{sam}^2 \leq \sigma_{all}^2$, através da obtenção de um intervalo de confiança unilateral de 95% para σ_{sam}^2 e rejeitando H quando este intervalo não incluir σ_{sam}^2. Isto equivale a rejeitar H quando $s_{sam}^2 > F_1 \sigma_{all}^2 + F_2 s_{an}^2$, onde s_{sam}^2 e s_{an}^2 são as estimativas usuais das variâncias amostral e analítica obtidas pela ANOVA, e F_1 e F_2 são constantes que podem ser derivadas de tabelas estatísticas padrão.

3.11.3 Tratamento de resultados dispersos em ensaios de homogeneidade

Dispersos analíticos esporádicos afetam os conjuntos de dados de ensaios de homogeneidade freqüentemente, já que pelo menos 20 resultados analíticos são produzidos em cada ensaio. Os dispersos analíticos manifestam-se como um grande e inesperado desvio entre os resultados duplicados em uma das amostras. Independentemente da heterogeneidade ou não do lote de material original, como o procedimento requer que cada amostra seja devidamente homogeneizada antes que as duas porções de ensaio sejam dela extraídas, qualquer diferença discrepante entre pares duplicados deve ser causada pela análise, mais do que pelo material .

O efeito de um resultado disperso (isto é, analítico) individual é talvez inesperado: embora inflacione a estimativa da variância interamostras, um disperso ajuda o material a ser aprovado no ensaio-*F* porque ele inflaciona a estimativa da variância analítica num grau proporcionalmente maior. Quanto mais extremo o disperso analítico, mais próximo se torna o valor-*F* do todo (outros resultados mantendo-se iguais). Assim, embora o Protocolo Harmonizado de 1993 determinasse que todos os resultados deveriam ser guardados, há uma clara razão para excluir dispersos analíticos extremos quando eles podem ser inequivocamente identificados. Entretanto, se um conjunto de dados aparentemente contém mais de um par de resultados analíticos discordantes, a validade de todo trabalho é colocada em dúvida e convém que os dados do ensaio de homogeneidade sejam descartados.

Note-se que a recomendação de rejeitar um par disperso individual apenas se aplica a amostras com resultados dispersos individuais, não a amostras com resultados mutuamente consistentes porém com médias dispersas. Se os resultados de uma amostra coincidem com os de outra mas o resultado médio é discordante com os outros dados, os resultados devem ser guardados—eles contêm evidências de heterogeneidade interamostras. Esta diferenciação é ilustrada na Figura 1. A Amostra 9 apresentou resultados discordantes que convém que sejam excluídos. A Amostra 12 apresentou resultados concordantes com uma média dispersa, e não convém que sejam excluídos. O teste de variância de Cochran é adequado para detectar diferenças extremas entre observações. (Apêndice 1).

Nota 1: A rejeição automática de valores dispersos de variância com o nível de confiança de 95% aumentará substancialmente a proporção de falsas reprovações em ensaios de homogeneidade e portanto não é recomendada.

Nota 2: Em circunstâncias raras, a recomendação de descartar um par individual de dispersos analíticos pode ser inadequada. Esta exceção pode ocorrer quando o analito está presente no material de ensaio em baixa concentração global, mas está quase que totalmente confinado a um traço que resiste à pulverização mas contém o analito em alta concentração. Um exemplo é uma rocha contendo ouro. A questão de se convém ou não usar a rejeição de dispersos é assunto para um comitê assessor do programa definir, levando em consideração a discussão acima.

Recomendação 9

Ao se ensaiar a homogeneidade suficiente, convém que os resultados duplicados de uma unidade de distribuição individual sejam descartados antes da análise da variância, caso sejam substancialmente diferentes um do outro no teste de Cochran a um nível de confiança de 99 % ou teste equivalente de variância intragrupos extrema. É recomendado que conjuntos de dados contendo discrepâncias em duas de tais unidades de distribuição sejam descartados totalmente. Convém que não sejam descartados pares de resultados com um valor de média disperso, mas sem evidência de variância extrema.

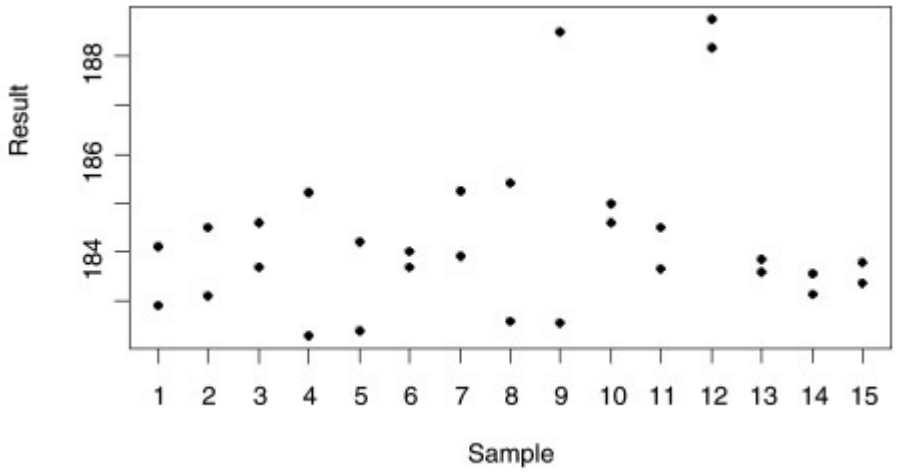


Fig. 1 Dados de ensaio de homogeneidade.

3.11.4 Outras anormalidades dos conjuntos de dados de ensaios de homogeneidade

Todos aspectos dos ensaios de homogeneidade suficiente dependem de que o laboratório implemente o ensaio corretamente e, especialmente, selecione as amostras de ensaio de maneira aleatória, homogeneizando-as antes da análise, analise as porções de ensaio duplicadas sob condições estritamente aleatórias, e registre os resultados com suficiente resolução numérica a fim de permitir a análise da variação. Qualquer infração a estas regras pode invalidar o resultado do ensaio. A não ser que um estrito controle seja mantido, encontra-se com freqüência conjunto de dados onde pelo menos alguns destes requisitos não foram observados. Portanto, recomenda-se que (a) instruções detalhadas sejam fornecidas ao laboratório que conduz o ensaio de homogeneidade, e (b) que os dados sejam verificados quanto a aspectos duvidosos de forma rotineira. No Apêndice 1 podem ser encontradas sugestões para estas instruções e ensaios.

Recomendação 10

a. Convém que instruções detalhadas sejam fornecidas ao laboratório que conduz o ensaio de homogeneidade.
b. Convém que os dados resultantes sejam verificados quanto a aspectos duvidosos.

3.11.5 Estabilidade de materiais de ensaio

Os materiais distribuídos em ensaios de proficiência devem ser suficientemente estáveis no período durante o qual o valor designado tenha que ser válido. O termo "suficientemente estável" implica em que quaisquer mudanças que ocorram durante o período pertinente devem ser de magnitude irrelevante em relação à interpretação dos resultados de uma rodada. Assim, caso seja estimado que uma mudança no índice-z (*z-score*) de ±1 seria irrelevante, então uma instabilidade que leve a uma mudança na concentração do analito da ordem de $0{,}1\,\sigma_p$ poderia ser aceitável. Normalmente, o período em questão é o intervalo entre a preparação do material e o prazo para o retorno dos resultados, embora o período possa ser mais longo caso material residual tenha que ser re-usado em rodadas subseqüentes ou para outros fins. Convém que o ensaio de estabilidade envolva exposição às condições mais extremas que possam ser encontradas durante a distribuição e armazenamento do material, ou a condições de degradação acelerada. Convém que o material sob ensaio seja acondicionado na embalagem na qual será distribuído.

Um ensaio abrangente de estabilidade suficiente demandaria recursos extremos (ver abaixo). Não é, portanto, praticável ensaiar todo lote de material para cada rodada de uma série. Entretanto, é uma precaução sensata ensaiar cada nova combinação de material e analito antes que a mesma seja usada pela primeira vez num ensaio de proficiência e eventualmente em outras, e os próximos parágrafos discutem isso. Pode também ser útil monitorar a estabilidade para, por exemplo, estabelecer que um laboratório individual seja encarregado de analisar as unidades pré e pós-distribuição, providenciando o retorno de algumas unidades de distribuição para comparação direta com unidades armazenadas, ou comparando resultados de análise pós-distribuição com informação prévia, tal como dados de ensaio de homogeneidade.

Ensaios básicos de estabilidade envolvem a comparação de níveis aparentes de analito entre materiais submetidos a condições de provável decomposição e materiais não submetidos a tais condições. Isto geralmente requer que uma amostra das unidades de distribuição seja aleatoriamente dividida em (pelo menos) dois subconjuntos iguais. O subconjunto "experimental" é submetido ao tratamento apropriado, enquanto que o subconjunto de "controle" é mantido em condições de máxima estabilidade, por exemplo, a baixas temperaturas e baixa pressão de oxigênio. Alternativamente, e particularmente se a estabilidade por períodos extensos é de interesse, o subconjunto de controle pode ser mantido em condições ambientais enquanto o subconjunto experimental é mantido sob condições de decomposição acelerada (ex: altas temperaturas). Os materiais são então analisados simultaneamente, ou se isso for impossível, como um delineamento em blocos aleatórios.

Tais experimentos devem ser cuidadosamente concebidos para evitar a combinação de efeitos de mudança planejados no material com variações na eficácia do método analítico usado. Análise do material de controle no começo do período de ensaio e do material experimental ao final automaticamente incluem qualquer diferença analítica de corrida a corrida nos resultados. Isto pode muito bem levar à conclusão incorreta de que há uma instabilidade significativa. A abordagem recomendada é, se possível, analisar ambos os subconjuntos (experimental e de controle) em ordem aleatória dentro de cada corrida de análise, isto é, sob condições de repetitividade. Qualquer diferença altamente significativa entre os resultados médios destes dois subconjuntos pode então seguramente ser considerado como evidência de instabilidade.

Como no ensaio de homogeneidade, uma diferenciação conceitual deve ser feita entre instabilidade estatisticamente significativa e instabilidade relevante. Por exemplo, uma mudança altamente significativa nos resultados analíticos pode ser detectada mas a mudança pode ainda ser tão pequena que apenas se pode inferir um efeito insignificante sobre índices-z (*z-scores*) dos participantes. Na prática, testes de significância não têm capacidade de validar instabilidade tão pequena, a não ser que seja usado um método analítico excepcionalmente preciso e/ou que se analise quantidades exageradas de unidades de distribuição. Um ensaio de estabilidade, portanto, apenas detectará uma instabilidade grosseira.

RECOMENDAÇÕES

Recomendação 1: Esquema para se obter um valor designado x e sua incerteza (ver Seção 3.3.2).

a. Excluir dos dados quaisquer resultados que sejam identificáveis como não válidos (exemplo: se expressos na unidade errada ou obtidos através do uso de um método proscrito) ou que sejam pontos dispersos extremos (exemplo: fora da faixa de ±50 % da mediana).

b. Examinar a disposição visual dos resultados remanescentes, através de um gráfico de pontos [para conjuntos pequenos de dados ($n < 50$)] ou através de um histograma (para conjuntos de dados maiores). Se os pontos dispersos fazem com que a disposição (gráfica) da maioria dos resultados se torne indevidamente comprimida fazer novo gráfico excluindo os pontos dispersos. Se a distribuição, pontos dispersos à parte, for

aparentemente unimodal e aproximadamente simétrica, ir para (c), senão ir para (d).

c. Calcular a média robusta $\hat{m}_{rob}$ e o desvio-padrão $\hat{s}_{rob}$ dos n resultados. Se $\hat{s}_{rob}$ for menor que aproximadamente $1{,}2\,s_p$, usar então $\hat{m}_{rob}$ como o valor designado x_a e $\hat{s}_{rob}/\sqrt{n}$ como sua incerteza-padrão. Se $\hat{s}_{rob} > 1{,}2\,s_p$, ir para (d).

d. Fazer uma estimativa da densidade *Kernel* da distribuição dos resultados usando *Kernel* normal com uma faixa h de $0{,}75\,s_p$. Se isto resultar numa densidade *Kernel* unimodal e aproximadamente simétrica, e a moda e a mediana forem praticamente coincidentes, usar então $\hat{m}_{rob}$ como o valor designado x_a e $\hat{s}_{rob}/\sqrt{n}$ como sua incerteza-padrão. Senão, ir para (e).

e. Se as modas secundárias podem ser atribuídas a resultados de pontos dispersos, e contribuem com menos de 5% da área total, então usar ainda $\hat{m}_{rob}$ como o valor designado x_a e $\hat{s}_{rob}/\sqrt{n}$ como sua incerteza-padrão. Senão, ir para (f).

f. Se as modas secundárias contribuem consideravelmente para com a área de *Kernel*, considerar a possibilidade de que duas ou mais populações discrepantes estejam representadas nos resultados dos participantes. Se for possível inferir, a partir de informação independente (ex: detalhes dos métodos analíticos dos participantes), que uma destas modas está correta e as outras estão incorretas, usar a moda selecionada como o valor designado x_a e seu desvio-padrão da média como sua incerteza-padrão. Senão, ir para (g).

g. Se os métodos acima falharem, abandonar a tentativa de determinar um valor de consenso e relatar que o desempenho de nenhum laboratório individual fez índices *(scores)* para aquela rodada. Entretanto, ainda pode ser útil fornecer aos participantes o sumário estatístico do conjunto de dados como um todo.

Recomendação 2: Uso da incerteza no valor designado (ver Seção 3.4)
Convém que o provedor de ensaios de proficiência indique um multiplicador $0{,}1 < l < 0{,}5$ apropriado para o ensaio, e tendo estimado $u^2(x_a)$ para uma rodada, agir do seguinte modo:

- se $u^2(x_a)/s_p^2 \le 0{,}1$, divulgar índices-z (*z-scores*) como desqualificado;
- se $0{,}1 < u^2(x_a)/s_p^2 \le l$, divulgar índices-z (*z-scores*) como qualificado, com observações (tais como "índices-z (*z-scores*) provisórios"); e
- se $u^2(x_a)/s_p^2 > l$ não divulgar índices-z (*z-scores*).

Nota: Na desigualdade $0{,}1 < l < 0{,}5$, os limites podem ser ligeiramente modificados a fim de atender a requisitos exatos de programas específicos.

Recomendação 3: Determinação do desvio-padrão para avaliação da proficiência (ver Seção 3.5) .Sempre que possível, o programa de ensaios de proficiência deve usar para s_p, o desvio-padrão para avaliação da proficiência, um valor que reflita a adequação aos propósitos para o setor. Se não existe um nível único que seja adequado de forma geral, convém que o provedor se abstenha de calcular índices *(scores)*, ou mostre claramente nos relatórios que os índices *(scores)* são para uso descritivo informal apenas e não para serem tomados como um índice de desempenho dos participantes.

Recomendação 4: O uso de ponderação no cálculo dos valores de consenso (ver Seção 3.6)
Mesmo quando as estimativas de incerteza são disponíveis, convém que se use métodos robustos não ponderados (isto é, que não consideram as incertezas individuais) para obter o valor de consenso e sua incerteza, conforme os métodos descritos nas Seções 3.3 e 3.4.

Recomendação 5: Atribuição de índices *(scores)* a resultados relatados com incerteza (ver Seção 3.6.2)
É recomendado que programas não forneçam índices *zeta*, a não ser que haja razões especiais para isso. Quando um participante apresenta requisitos inconsistentes com aqueles do programa, o participante pode calcular índices *zeta* ou equivalentes.

Recomendação 6: Classificação e ordenação de laboratórios (ver Seção 3.9)
Os provedores de programas de proficiência, participantes e usuários finais devem evitar a classificação e ordenação de laboratórios com base em seus índices-z (*z-scores*).

Recomendação 7: Requisito de repetitividade em ensaios de homogeneidade (ver Seção 3.11.1)
A precisão analítica (repetitividade) do método usado no ensaio de homogeneidade deve satisfazer $\sigma_{an} / \sigma_p < 0{,}5$ onde σ_{an} é o desvio-padrão da repetitividade adequado ao ensaio de homogeneidade.

Recomendação 8: Teste estatístico em ensaios de homogeneidade (ver Seção 3.11.2)
Empregar um teste explícito de hipótese H: $\sigma^2_{sam} \leq \sigma^2_{all}$, encontrando um intervalo de confiança unilateral de 95% para σ^2_{sam} e rejeitando H quando este intervalo não incluir σ^2_{all} . Isto eqüivale a rejeitar H quando

$$s^2_{sam} > F_1 \sigma^2_{all} + F_2 s^2_{an} \quad (6)$$

onde s^2_{sam} e s^2_{an} são estimativas usuais das variâncias amostral (s^2_{sam}) e analítica (s^2_{an}) obtidas da ANOVA, e F_1 e F_2 são constantes que podem ser derivadas de tabelas estatísticas padrão.

Recomendação 9: Tratamento de pontos dispersos em ensaios de homogeneidade (ver Seção 3.11.3)
Ao se ensaiar para suficiente homogeneidade, os resultados duplicados de uma unidade de distribuição individual devem ser descartados antes da análise da variância, caso sejam significativamente diferentes um do outro no teste de Cochran a um nível de confiança de 99 % ou num teste equivalente para extrema variância intra-grupo. É recomendado que conjuntos de dados contendo discrepâncias em duas de tais distribuições sejam totalmente descartados. Pares de resultados com um valor de média disperso mas sem evidência de variância extrema não devem ser descartados.

Recomendação 10: Condução de ensaios de homogeneidade (ver Seção 3.11.4)
a. Instruções detalhadas devem ser fornecidas ao laboratório que conduz o ensaio de homogeneidade.
b. Os dados resultantes devem ser verificados quanto a características duvidosas.

REFERÊNCIAS

1. M. Thompson and R. Wood. "The International Harmonised Protocol for the proficiency testing of (chemical) analytical laboratories", *Pure Appl. Chem.* **65**, 2123-2144 (1993). [Também publicado *no J. AOAC Int*. **76**, 926-940 (1993)].

2. Ver: (a) M. Golze. "Information system and qualifying criteria for proficiency testing schemes", *Accred. Qual. Assur*. 6, 199-202 (2001); (b) J. M. F. Nogueira, C. A. Nieto-de-Castro, L. Cortez. "EPTIS: The new European database of proficiency testing schemes for analytical laboratories", *J. Trends Anal. Chem.* 20, 457-61 (2001); (c) <http://www.eptis.bam.de>.

3. R. E. Lawn, M. Thompson, R. F. Walker. Proficiency Testing in Analytical Chemistry, The Royal Society of Chemistry, Cambridge (1997).

4.International Organization for Standardization. ISO Guide 43: Proficiency testing by interlaboratory comparisons—Part 1: Development and operation of proficiency testing schemes, Geneva, Switzerland (1994)

5.International Organization for Standardization. ISO 13528: Statistical methods for use in proficiency testing by interlaboratory comparisons, Geneva, Switzerland (2005).

6.ILAC-G13:2000. Guidelines for the requirements for the competence of providers of proficiencytesting schemes. Available online at <http://www.ilac.org/>.

7.M. Thompson, S. L. R. Ellison, R. Wood. "Harmonized guidelines for single laboratory validation of methods of analysis", Pure Appl. Chem.74, 835–855 (2002).

8.International Organization for Standardization. ISO Guide 33:2000, Uses of Certified Reference

Materials, Geneva, Switzerland (2000).

9.M. Thompson and R. Wood. “Harmonised guidelines for internal quality control in analytical chemistry laboratories”, Pure Appl. Chem.67, 649–666 (1995).

10.T. Fearn, S. Fisher, M. Thompson, S. L. R. Ellison. “A decision theory approach to fitness-for-purpose in analytical measurement”, Analyst127, 818–824 (2002).

11.Analytical Methods Committee. “Robust statistics—how not to reject outliers: Part 1 Basic con-cepts”, Analyst114, 1693 (1989).

12.M. Thompson. “Bump-hunting for the proficiency tester: Searching for multimodality”, Analyst127,1359–1364 (2002).

13.M. Thompson. “A natural history of analytical methods”, Analyst124, 991 (1999).

14.M. Thompson. “Recent trends in interlaboratory precision at ppb and sub-ppb concentrations inrelation to fitness-for-purpose criteria in proficiency testing”, Analyst125, 385–386 (2000).

15.Analytical Methods Committee. “Uncertainty of measurement—implications for its use in ana-lytical science”, Analyst120, 2303–2308 (1995).

16.W. P. Cofino, D. E. Wells, F. Ariese, J.-W. M. Wegener, R. I. H. M. Stokkum, A. L. Peerboom. “A new model for the inference of population characteristics from experimental data using uncertainties. Application to interlaboratory studies”, Chemom. Intell. Lab Systems53, 37–55 (2000).

17.T. Fearn. “Comments on ‘Cofino Statistics’”, Accred. Qual. Assur. 9, 441–444 (2004).

18.M. Thompson and P. J. Lowthian. “The frequency of rounds in a proficiency test: does it affectthe performance of participants?”, Analyst123, 2809–2812 (1998).

19.T. Fearn and M. Thompson. “A new test for sufficient homogeneity”, Analyst 126, 1414–1417(2001).

APÊNDICE 1: PROCEDIMENTO RECOMENDADO PARA ENSAIAR MATERIAL QUANTO À HOMOGENEIDADE SUFICIENTE

A1.1 Procedimento

1. Preparar todo o lote de material numa forma que seja considerada homogênea, por meio de método apropriado.
2. Dividir o material nos recipientes que serão usados para envio aos participantes.
3. Selecionar um mínimo de 10 recipientes de maneira rigorosamente aleatória.
4. Homogeneizar separadamente o conteúdo de cada um dos *m* recipientes selecionados e extrair duas porções de cada.
5. Analisar as 2m porções de ensaio em ordem aleatória sob condições de repetitividade através de um método apropriado. O método analítico usado deve ser suficientemente preciso para permitir uma estimativa satisfatória de s_{sam}. Se possível, $\sigma_{an} < 0{,}5\,\sigma_p$.

O primeiro passo é verificar os dados quanto a patologias. Tal verificação poderia ser feita visualmente num gráfico simples de resultados × número de amostra, procurando por características usadas em diagnóstico tais como: (a) tendências ou descontinuidades, (b) distribuição não-aleatória de diferenças entre o primeiro e segundo resultado de ensaio, (c) arredondamento excessivo; e (d) resultados dispersos intra-amostras.

Se tudo estiver bem, usar os dados para estimar as variâncias analítica e amostral. Caso um programa para executar análise de variância fator único estiver disponível, pode ser usado. Alternativamente, um esquema de cálculo completo é apresentado abaixo.

A1.2 Métodos estatísticos

A1.2.1 Procedimento do teste de Cochran para resultados duplicados

Calcular a soma, S_i, e a diferença, D_i, de cada par de duplicatas, para i = 1, ..., m.

Calcular a soma dos quadrados S_{DD} das m diferenças a partir da equação:

$$S_{DD} = \sum_m D_i^2$$

A estatística do teste de Cochran é a razão entre $D_{\max}^2$, a maior diferença ao quadrado e a soma dos quadrados das diferenças:

$$C = \frac{D_{\max}^2}{S_{DD}}$$

Calcular a razão acima e comparar com os valores críticos apropriados das tabelas. A Tabela 1 mostra os valores críticos para 95 e 99% de confiança para m entre 7 e 20 pares.

Os resultados para pares dispersos de Cochran detectados para 95% ou mais de nível de confiança devem ser sempre verificados quanto a evidências de erros de transcrição ou outros erros na análise, e convém que sejam implementadas medidas apropriadas caso quaisquer erros sejam detectados. Convém que um par disperso não seja rejeitado a não ser que seja significativo ao nível de 99% ou sejam detectados erros insanáveis de procedimento analítico. Convém que um único valor disperso de Cochran para um nível de 99% seja excluído da ANOVA, a não ser que haja razão para o contrário (ver Seção 3.11).

A1.2.2 Ensaio de não-homogeneidade significativa

Usar a mesma soma dos quadrados das diferenças para calcular

$$s_{an}^2 = \sum D_i^2 / 2m$$

Calcular a variância V_S das somas S_i

$V_S = \sum \frac{(s_i - \bar{s})^2}{(m-1)}$, onde $\bar{s} = (1/m)\sum s_i$ é a média de s_i.

Calcular a variância amostral s^2_{sam} como:

$s^2_{sam} = \frac{(V_S / 2 - s_{an})}{2}$, ou como $s^2_{sam} = 0$ se a estimativa acima for negativa.

Se um programa de análise da variância fator único estiver disponível, as grandezas $V_S / 2$ e s_{an} acima podem ser extraídas da tabela de análise da variância como os quadrados médios "entre" e "dentro", respectivamente.

Calcular a variância amostral aceitável σ^2_{all} como

$\sigma^2_{all} = (0{,}3\sigma_p)^2$, onde σ_p é o desvio-padrão para avaliação da proficiência.

Tomando-se os valores de F_1 e F_2 da Tabela 2, calcular o valor crítico para o ensaio como:

$c = F_1\sigma^2_{all} + F_2 s^2_{an}$.

Se $s^2_{sam} > c$, há evidência (significativo para o nível de confiança de 95 % de confiança) de que o desvio-padrão amostral na população de amostras ultrapassa a fração aceitável do desvio-padrão alvo, e o teste para homogeneidade foi reprovado .

Se $s^2_{sam} < c$, não há tal evidência, o teste de homogeneidade foi aprovado.

A1.2.3 Tabelas de valores críticos para teste de homogeneidade

Tabela 1 Valores críticos para o testede Cochran para duplicatas

m	95%	99%
7	0,727	0,838
8	0,680	0,794
9	0,638	0,754
10	0,602	0,718
11	0,570	0,684
12	0,541	0,653
13	0,515	0,624
14	0,492	0,599
15	0,471	0,575
16	0,452	0,553
17	0,434	0,532
18	0,418	0,514
19	0,403	0,496
20	0,389	0,480

Tabela 2 Fatores F_1 e F_2 para uso em teste de homogeneidade suficiente.

*m**	20	19	18	17	16	15	14	13	12	11	10	9	8	7
F_1	1,59	1,60	1,62	1,64	1,67	1,69	1,72	1,75	1,79	1,83	1,88	1,94	2,01	2,10
F_2	0,57	0,59	0,62	0,64	0,68	0,71	0,75	0,80	0,86	0,93	1,01	1,11	1,25	1,43

m é o número de amostras que foram medidas em duplicata.*

As duas constantes na Tabela 2 são derivadas das tabelas de estatísticas padrão como $F_1 = \frac{\chi^2_{m-1,0.95}}{(m-1)}$, onde $\chi^2_{m-1,0.95}$ é o valor ultrapassado com probabilidade de 0,05 por uma variável aleatória qui-quadrado com

$m-1$ graus de liberdade, e $F_2 = \frac{(F_{m-1,m,0.95}-1)}{2}$, onde $F_{m-1,m,0.95}$ é o valor ultrapassado com probabilidade de 0,05 por uma variável aleatória com uma distribuição-*F* com *m*-1 e *m* graus de liberdade.

A1.3 Exemplos de instruções para laboratórios que executam ensaios de homogeneidade suficiente

Convém que o laboratório seja experiente no método analítico utilizado.

- Selecionar as $m \geq 10$ unidades de distribuição, de forma rigorosamente aleatória, a partir do conjunto completo de unidades. Isto deve ser feito de maneira formal, designando um número seqüencial para as unidades, seja explicitamente (rotulando-as) ou implicitamente (por exemplo: sua posição num arranjo). Fazer a seleção usando números aleatórios de uma tabela ou gerados (reiniciando a cada vez) por um programa de software (ex: Microsoft Excel). Não é aceitável selecionar as unidades de qualquer outra forma (por exemplo, misturando-as). Não utilizar uma seqüência aleatória previamente usada.
- Homogeneizar os conteúdos da cada unidade de distribuição de uma forma adequada (ex: num misturador) e retirar de cada uma delas duas porções. Rotular as porções de ensaio conforme mostrado.

Código seqüencial da unidade de distribuição	Rótulo da primeira porção do ensaio	Rótulo da segunda porção do ensaio
1	1.1	1.2
2	2.1	2.2
3	3.1	3.2
.	.	.
.	.	.
m	*m*.1	*m*.2

- Arranjar as 20 porções de ensaio em ordem aleatória e implementar todas as operações analíticas nelas segundo aquela ordem. Novamente, tabelas de números aleatórios ou um pacote de software devem ser usados para gerar uma nova seqüência aleatória. Um exemplo de seqüência aleatória (não deve ser copiado) é 7,1 3,1 5,2 5,1 10,2 1,1... .2,1 9,2 8,2 1,2 4,1 2,2 9,1 10,1 7,2 3,2 8,1 6,1 4,2 6,2.
- Conduzir a análise preferivelmente sob condições de repetitividade (isto é., em uma corrida) ou, se isso for impossível, em corridas sucessivas com o menor número possível de mudanças no sistema analítico, usando um método que tenha um desvio-padrão da repetitividade menor que 0,5 σ_p. Registrar os resultados se possível com tantos algarismos significativos quanto requerido por 0,01 σ_p.

- Retornar os 20 resultados analíticos, incluindo os rótulos, na ordem usada na corrida.

A1.4 Ensaio de homogeneidade: Exemplo

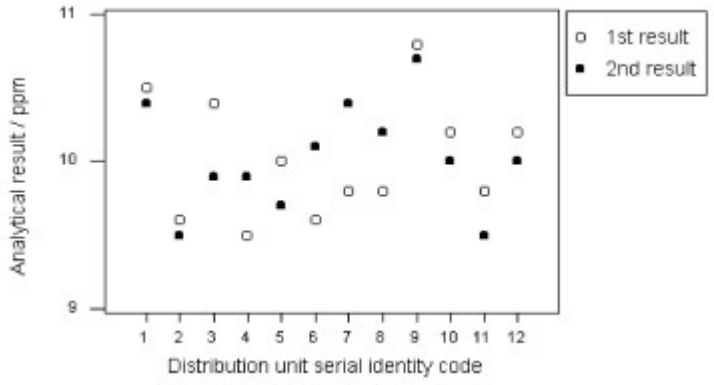
A1.4.1 Os dados

Os dados, mostrados na Tabela 3, são extraídos do Protocolo Harmonizado de 1993 [Ref. 1 da Parte 3].

Tabela 3 Resultados de duplicatas para 12 unidades de distribuição de farinha de soja analisados para detecção de cobre (ppm), junto com alguns estágios intermediários do cálculo da ANOVA.

Amostra	Resultado a	Resultado b	$D = a - b$	$S = a + b$	$D^2 = (a-b)^2$
1	10,5	10,4	0,1	20,9	0,01
2	9,6	9,5	0,1	19,1	0,01
3	10,4	9,9	0,5	20,3	0,25
4	9,5	9,9	-0,4	19,4	0,16
5	10,0	9,7	0,3	19,7	0,09
6	9,6	10,1	-0,5	19,7	0,25
7	9,8	10,4	-0,6	20,2	0,36
8	9,8	10,2	-0,4	20,0	0,16
9	10,8	10,7	0,1	21,5	0,01
10	10,2	10,0	0,2	20,2	0,04
11	9,8	9,5	0,3	19,3	0,09
12	10,2	10,0	0,2	20,2	0,04

A1.4.2 Apreciação visual



Os dados são visualmente apresentados acima, e não mostram nenhum aspecto duvidoso tal como resultados duplicados discordantes, amostras dispersas, tendências, descontinuidades ou qualquer outro efeito sistemático. (Um gráfico de Youden da primeira versus a segunda duplicata também poderia ser usado).

A1.4.3 Teste de Cochran

O maior valor de D^2 é 0,36 e a soma de D^2 é 1,47, assim , o resultado do teste de Cochran is 0,36/1,47 = 0,24. Isto é menor do que o valor crítico (5%) de 0,54 , assim não há evidência de dispersos analíticos e prossegue-se com o conjunto de dados completo.

A1.4.4 Teste de homogeneidade

Variância analítica: s^2_{an} =1,47/24 = 0,061 .

Variância entre amostras: A variância da soma $S = a + b$ é 0,463 , assim

s^2_{sam} = (0,463/2 – 0,061)/2 = (0,231 – 0,061)/2 = 0,085

Variância entre amostras aceitável: o desvio-padrão alvo é 1,14 ppm, assim a variância entre amostras aceitável é $\sigma_{all}^2 = (0{,}3 \times 1{,}14)^2 = 0{,}116$.

Valor crítico: O valor crítico para o ensaio é $1{,}79\ \sigma_{all}^2 + 0{,}86\ s_{an}^2 = 1{,}79 \times 0{,}116 + 0{,}86 \times 0{,}061 = 0{,}26$

Uma vez que $s_{sam}^2 = 0{,}085 < 0{,}26$, houve aprovação do teste e o material é suficientemente homogêneo.

APÊNDICE 2: EXEMPLO DE COMO CONDUZIR UM ENSAIO DE ESTABILIDADE

O procedimento descrito na Seção 3.11.5 foi implementado. O desvio-padrão para avaliação da proficiência (σ_p) foi estabelecido em 0,1*c* (isto é, um desvio-padrão relativo (DPR) de 10 %) e os resultados da análise sob condições de repetitividade, na ordem aleatória na qual as análises foram feitas, estão tabuladas abaixo).

Material	Resultado / ppm
Experimental	11,5
Controle	13,4
Controle	12,2
Experimental	12,3
Controle	12,7
Experimental	10,9
Controle	12,5
Experimental	11,4
Experimental	12,4
Controle	12,5

Um teste-*t* de duas amostras com desvio-padrão combinado dá as seguintes estatísticas:

	n	$\bar{x}$
Controle	5	12,66
Experimental	5	11,70
Diferença		0,96

Desvio-padrão combinado: 0,551
Intervalo de confiança de 95 % para ($\mu_{cont} - \mu_{exp\,t}$): (0,16; 1,76)

O teste-*t* de $H_0: \mu_{cont} = \mu_{exp\,t}$ contra $H_A: \mu_{cont} \neq \mu_{exp\,t}$ dá um valor de t = 2,75 com 8 graus de liberdade, correspondendo a uma probabilidade (valor-*p*) de 0,025. A diferença de instabilidade de 0,96 ppm é, portanto, significante ao nível de confiança de 95%. (Isto também pode ser deduzido do intervalo de confiança de 95%, que não inclui o zero).

Utilizando-se a média de aprox. 12 como a concentração do analito, achamos $\sigma_p = 0{,}1 \times 12 = 1{,}2$. A diferença em função da instabilidade é muito maior que o limite desejado de $0{,}1\ \sigma_p$, portanto há instabilidade relevante e o material é inapropriado para uso.

APÊNDICE 3: EXEMPLOS DA PRÁTICA NA DETERMINAÇÃO DE UM CONSENSO DE PARTICIPANTES PARA USO COMO VALOR DESIGNADO

A3.1 Exemplo 1

Os resultados dos participantes são listados na tabela seguinte, a qual também mostra o resumo estatístico pertinente. A unidade é fração de massa, expressas como porcentagem (%); a precisão numérica é conforme relatada pelos participantes.

Resultados relatados

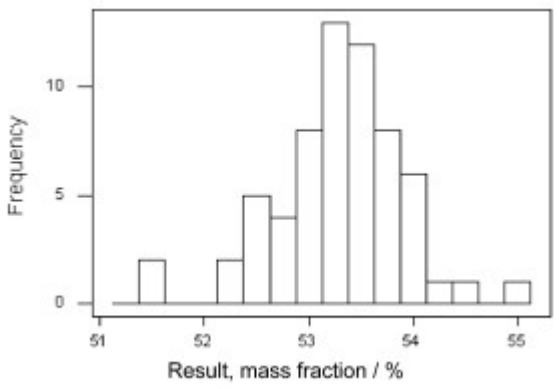
54,09	53,15	53,702	52,9	53,65	52,815	53,5
52,95	52,35	53,49	55,02	53,32	54,04	53,15
53,41	53,4	53,3	54,33	52,83	53,4	53,38
53,19	52,4	52,9	53,44	53,75	53,39	53,661
54,09	53,09	53,21	53,12	53,18	53,3	52,62
53,7	53,51	53,294	53,57	52,44	53,04	53,23
63,54	46,1	53,18	54,54	53,76	54,04	53,64
53	54,1	52,2	52,54	53,42	53,952	50,09
53,06	48,07	52,51	51,44	52,72	53,7	
53,16	53,54	53,37	51,52	46,85	52,68	

Resumo estatístico

n	68
Média	53,1
Desvio-padrão	1,96
Mediana	53,3
Estimativa H15 da média	53,24
Estimativa H15 do desvio-padrão*	0,64

*Ver refs. [1] e [2] para este Apêndice.

O desvio-padrão para a avaliação da proficiência s_p é 0.6 %



O histograma (valores dispersos à parte) sugere uma distribuição unimodal e aproximadamente simétrica.

O resumo estatístico mostra uma média e mediana robustas quase coincidentes. O desvio-padrão robusto é menor que 1,2 σ_p, assim não há preocupações quanto a distribuições amplas.

O valor de $\hat{s}_{rob}/\sqrt{n}$ = 0,079 , está bem abaixo do parâmetro de 0,3 σ_p = 0,17.

O valor de consenso e sua incerteza-padrão são obtidos como frações de 53,24 e 0,08% de massa, respectivamente.

A3.2 Exemplo 2

Os participantes relataram os seguintes resultados (unidades são ppb, isto é, 10^{-9} fração de massa)

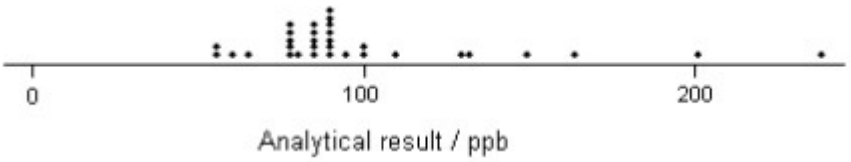
Resultados relatados/ppb

133	89	55	84,48	84,4	90,4	66,6
77	80	60,3	84	78	85	130
90	79	99,7	149	91	164	
78	84	110	77	91	89	
95	55	90	100	200,56	237	

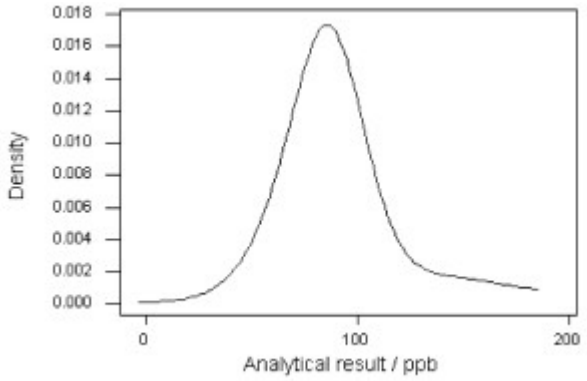
Resumo estatístico

n	32
Média	99,26
Desvio-padrão	39,76
Mediana	89
Estimativa H15 da média	91,45
Estimativa H15 do desvio-padrão*	23,64

Um gráfico de pontos (abaixo) dos resultados relatados mostra um conjunto de dados com uma forte assimetria positiva, o que poderia causar dúvidas sobre a validade dos resultados estatísticos robustos.



Um desvio-padrão provisório para avaliação da proficiência, derivado da média robusta, foi obtido pela função de *Horwitz*: $\sigma_p = 0{,}452 \times 91{,}4^{0{,}8495} = 20{,}8$ ppb. Devido à assimetria e ao desvio-padrão robusto alto, a média robusta fica sob suspeita, de forma que uma distribuição de densidade de *kernel* foi construída com uma faixa h de 0,75 σ_p:



A densidade de *kernel* mostra uma moda de 85,2 ppb, e a reamostragem (o "bootstrapping") dos dados forneceu um erro padrão para a moda de 2,0 ppb. O σ_p revisado com base numa concentração de 85,2 é 19,7 ppb. A incerteza implícita da moda (2,0) está abaixo do parâmetro de 0,3 σ_p = 5,9 ppb.

O valor de consenso e sua incerteza-padrão obtidos são 85 e 2 ppb, respectivamente.

A3.3 Exemplo 3

Esta é a primeira rodada de uma série depois que o método de relato foi modificado, de forma a quantificar uma diferente "forma de ponderação". A razão entre as massas moleculares relativas das formas nova e antiga de ponderação foi 1,35.

O desvio-padrão para avaliação da proficiência para a série é determinado pela Função de *Horwitz* $\sigma_p = 0{,}16 \times c^{0{,}8495}$, onde σ_p e c estão em ppm (10^6 fração de massa).

Os resultados dos participantes, (em ppm, isto é, 10^6 fração de massa) foram:

Resultados relatados / ppm

102,5	97,9	102	101	99	75,9	101
74	94	93	70	82,9	106	113
122	97	114	101	70	103,88	93
96	107	103	96	119	99	83
107	101	134	109	103,8	106	77
95	108	96	104	101,33	92,2	
94,5	102	77	98,91	107	109	
89	110	103	112	55	87	
108	105,4	86	74	73	77	
96	77,37	73,5	78	92	84,6	

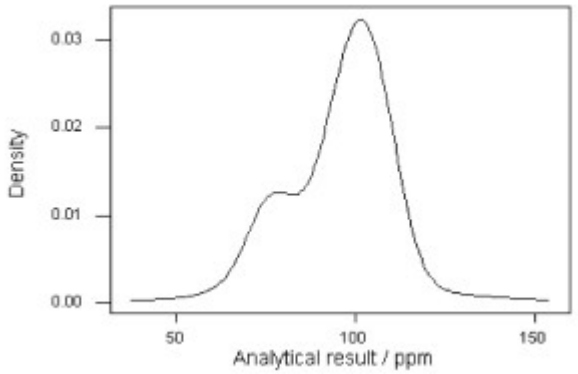
Resumo estatístico

n	65
Média	95,69
Desvio-padrão	14,52
Mediana	98,91
Estimativa H15 da média	95,78
Estimativa H15 do desvio-padrão	14,63

Conforme mostra o gráfico de pontos abaixo, isto possivelmente representa uma população bimodal.

O valor provisório s_p derivado da média robusta é 7,71 ppm, mas o desvio-padrão robusto é consideravelmente maior, assim há sólidos fundamentos para se suspeitar de uma distribuição mista.

Uma densidade de *kernel* foi obtida usando-se uma amplitude *h* de 0,75 × 7,71 = 5,78.



Esta função de densidade tem modas de 78,6 e 101,5 ppm, com erros padrão estimados pela reamostragem de 13,6 e 1,6 , respectivamente. A razão entre as modas é 101,5/78,6 = 1,29 , a qual se aproxima da razão de 1,35 das massas moleculares relativas, o que justifica a suposição de que a maior moda está correta e a menor moda representa os participantes que relataram incorretamente os resultados na forma antiga de ponderação.

O consenso, portanto, é identificado como 101,5 com uma incerteza de 1,6 ppm. O valor alvo revisto baseado neste consenso é 8,1. Como a incerteza de 1,6 é menor que 0,3 × 8,1 = 2,43, a incerteza é aceitável e o consenso pode ser usado como o valor designado.

REFERÊNCIAS

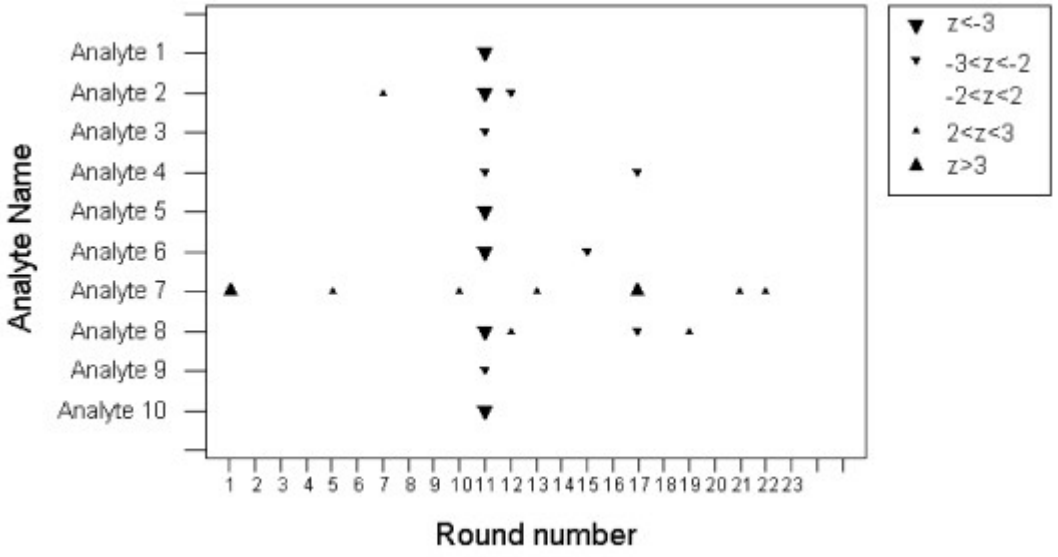
1. Analytical Methods Committee. "Robust statistics—how not to reject outliers: Part 1 Basic concepts", *Analyst* **114**, 1693 (1989).
2. International Organization for Standardization. *ISO 13528: Statistical methods for use in proficiency testing by interlaboratory comparisons*, Geneva, Switzerland (2005).

APÊNDICE 4: AVALIANDO ÍNDICES-Z (Z-SCORES) NO LONGO PRAZO: ÍNDICES *(SCORES)* SUMÁRIOS E MÉTODOS GRÁFICOS

Enquanto um índice-z (*z-score*) individual representa uma indicação útil do desempenho de um laboratório, a consideração de um conjunto ou seqüência de índices-z (*z-scores*) fornece uma visão mais profunda. Além disso, índices-z (*z-scores*) obtidos ao longo do tempo (para combinação específica de um analito/material de ensaio) refletem-se na incerteza do participante. Tanto os métodos gráficos quanto os estatísticos podem ser adequados para se examinar conjuntos de índices-z (*z-scores*). Entretanto, ao interpretar os sumários estatísticos, o devido cuidado é necessário para se evitar conclusões incorretas. É especialmente enfatizado que não é recomendado o uso de um índice sumário derivado de índices-z (*z-scores*) relativos a analitos diferentes; ele tem uma faixa muito limitada de aplicação válida e tende a mascarar problemas aleatórios ou sistemáticos em analitos individuais. Além disso, pode ser utilizado inadequadamente por não cientistas.

A4.1 Índices sumários

Os dois tipos seguintes de índices sumários são estatisticamente bem fundamentados e podem ser úteis para



z_n]

deriv

pode a em zero útil prop 1,5; 1,5; ados indiv inais opos

$$S_{ZZ} = \sum_i z_i^2$$

poderia ser interpretada como uma distribuição χ^2_n para índices-z (*z-scores*) centrados em zero com a unidade de variância. Esta estatística tem a vantagem de evitar o cancelamento de grandes índices-z (*z-scores*) de sinais opostos, mas é menos sensível a pequenas tendências (*bias*).

Ambos resumos estatísticos precisam ser protegidos (ex: filtrando-as ou tornando-as robustas) dos índices (*scores*) de dispersos passados, os quais, de outro modo, teriam uma persistência de longo prazo. S_{ZZ} é especialmente sensível a dispersos. Ambas estatísticas, (quando robustas) podem ser relacionadas à incerteza de medição da seguinte maneira. Se os índices-z (*z-scores*) são baseados em adequação aos propósitos e portanto assumidos como aleatórios $N(0,1)$, significativos níveis altos nas estatísticas indicam que a incerteza de medição dos participantes é maior que aquela indicada pelo critério de adequação aos propósitos do programa.

A4.2 Métodos Gráficos

Métodos gráficos conclusivos sobre um conjunto de índices-z (*z-scores*) podem ser tão informativos quanto os índices sumários e menos sujeitos à má-interpretação. Gráficos de Shewart (com limites de advertência e ação em $z = \pm 2$ e $z = \pm 3$, respectivamente) podem ser aplicados. Múltiplos gráficos simbólicos univariados [1], tais como os mostrados abaixo, dão uma visão clara e são especialmente úteis quando são considerados os índices *(scores)* de um grupo de analitos determinados por um método comum. Gráficos desenhados à mão são rapidamente atualizados e são tão adequados quanto os produzidos por computadores.

O gráfico de controle (Fig. A4.1) mostra símbolos apontando para cima para indicar índices-z (*z-scores*) maiores que zero e símbolos apontando para baixo para aqueles menores que zero. Pequenos símbolos representam situações onde $2 \leq |z| < 3$, e símbolos grandes situações onde $|z| \geq 3$. Os dados ilustrados imediatamente mostram algumas qualidades dignas de nota. Os resultados da rodada 11 foram na maioria muito baixos, indicando um procedimento que estava falho em relação a algum aspecto geral, enquanto analito 7 dá resultados altos com muita freqüência, demonstrando um problema persistente com aquele analito específico. Os resultados remanescentes são aproximadamente consistentes com a adequação aos propósitos, o que na média resultaria em aproximadamente 5% de índices-z (*z-scores*) representados por pequenos símbolos.

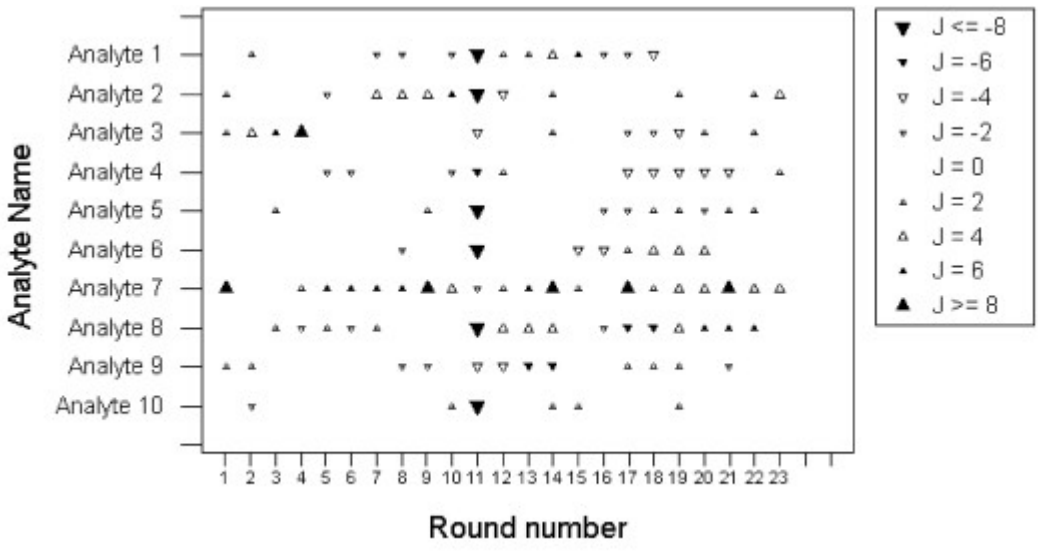


Fig A4.1 Gráfico de controle para índices-z (*z-scores*)

Um gráfico-*J* (também conhecido como "gráfico de zona") é ainda mais informativo, porque ele combina as capacidades dos gráficos de Shewart e dos gráficos CUSUM (Somas Acumuladas). Isto acontece através da acumulação de índices-J especiais atribuídos aos sucessivos resultados em qualquer lado do eixo do zero. Isto permite detectar tendências mínimas persistentes assim como grandes mudanças abruptas no sistema analítico.

As regras gerais para conversão de índices-z (*z-scores*) em índices-J são as seguintes:

Se $z \geq 3$	então $J = 8$.
Se $2 \leq z < 3$	então $J = 4$.
Se $1 \leq z < 2$	então $J = 2$.
Se $-1 < z < 1$	então $J = 0$.
Se $-2 < z \leq -1$	então $J = -2$.
Se $-3 < z \leq -2$	então $J = -4$.
Se $z \leq -3$	então $J = -8$.

Os índices-J de rodadas sucessivas são acumulados até que $|z| \geq 8$, o que significa uma incursão além dos limites de ação, e assim procedimentos investigativos são acionados. O acumulador é zerado imediatamente (isto é, antes de qualquer nova acumulação) depois de tal incursão e quando os sucessivos valores de J são de sinais opostos.

Vários exemplos de efeitos acumulados de tendências (*bias*) podem ser vistos na Fig. A4.2 (a qual ilustra os mesmos resultados como Fig. A4.1 para comparação). Por exemplo, o analito 3 nas rodadas de 1 a 4 recebe índices-z (*z-scores*) de 1,5; 1,2; 1,5; e 1,1 respectivamente, o que se traduz em índices-*J* de 2, 2, 2, e 2, o que leva ao acúmulo de 8 por volta da rodada 4 e aciona procedimentos investigativos. Exemplos similares podem ser vistos para o analito 7.

Fig. A4.2 Gráfico-J para índices-z (*z-scores*) (mesmos dados do gráfico anterior).

REFERÊNCIAS

1. M. Thompson, K. M. Malik, R. J. Howarth. "Multiple univariate symbolic control chart for internal quality control of analytical data", *Anal. Comm*. **35**, 205-208 (1998).
2. Analytical Methods Committee. "The J-chart: A simple plot that combines the capabilities of Shewhart and cusum charts, for use in analytical quality control", AMC Technical Briefs: No 12. <www.rsc.org/amc/>.

APÊNDICE 5: VALIDAÇÃO DE MÉTODOS ATRAVÉS DOS RESULTADOS DE ENSAIOS DE PROFICIÊNCIA

O propósito do programa de ensaios de proficiência é testar a exatidão dos laboratórios participantes. Os participantes têm a liberdade de escolher o método de análise e geralmente usam uma multiplicidade de métodos (ou variantes de um "método" individual). Consequentemente, não há geralmente possibilidade de validar métodos como um subproduto de ensaios de proficiência. Entretanto, a validação de métodos se torna uma possibilidade, caso haja número suficiente de participantes no programa de proficiência usando métodos de análise similares. Esta possibilidade, caso adequadamente explorada, pode ser vista como uma alternativa mais barata ao, ou uma confirmação de, estudo colaborativo, o qual é um modelo reconhecido (mas caro) para validação interlaboratorial de métodos. (Os estudos colaborativos custam normalmente € 30.000 para cada método).

Os programas de ensaios de proficiência, entretanto, diferem dos estudos colaborativos, em concepção e resultados, numa variedade de formas e conseqüências.

- Freqüentemente, apenas um material de ensaio (ou um pequeno número deles) é enviado em cada rodada, comparado com o mínimo de 5 em estudos colaborativos. É, portanto, necessário coletar dados de várias rodadas, por um período de talvez alguns anos, para se obter informação suficiente para propósitos de validação. (É importante lembrar neste contexto que, estritamente falando, não se valida "um método" como uma entidade isolada. O que se valida é um método aplicado a analitos específicos e faixas definidas de concentração de analitos e de materiais de ensaio. Assim, nem todas as rodadas numa série podem ser elegíveis para uso na validação).
- Os programas de ensaios de proficiência raramente demandam o relato de resultados duplicados, e assim as estimativas do desvio-padrão da repetitividade não estão disponíveis nos resultados dos ensaios de proficiência. (Isto não é uma grande perda—é fácil para os laboratórios estimar os seus próprios desvios-padrão da repetitividade).
- Em programas de ensaios de proficiência não há garantia de que os mesmos laboratórios participarão do programa em diferentes rodadas.
- Em um estudo colaborativo, os participantes são selecionados com base na competência provável. Em ensaios de proficiência, a competência universal não é uma suposição sensata.

Com a devida consideração a estas diferenças, os resultados dos ensaios de proficiência, restritos a participantes que usam um protocolo de método definido, podem ser usados para estimar, razoavelmente, o desvio-padrão da reprodutibilidade do método [2]. Para se alcançar o resultado desejado, requer-se a utilização de métodos de estimativa robustos combinados ao julgamento de especialistas. Se dois ou mais métodos definidos de forma estreitamente similar são usados por um número suficiente de participantes, é possível avaliar qualquer tendência entre os métodos ao longo de uma extensa faixa de concentrações [3,4] , através da estimativa da relação funcional [5,6].

REFERÊNCIAS

1. W. *Horwitz* (Ed.). "Protocol for the design, conduct and interpretation of method performance studies", *Pure Appl. Chem.* **67**, 331-343 (1995).
2. Paper CX/MAS 02/12 of the Codex Committee on Methods of Analysis and Sampling. Validation of Methods Through the Use of Results from Proficiency Testing Schemes, Twenty-fourth Session, Budapest, Hungary, 18-22 November 2002, FAO, Rome.
3. P. J. Lowthian, M. Thompson, R. Wood. "The use of proficiency tests to assess the comparative performance of analytical methods: The determination of fat in foodstuffs", *Analyst* **121**, 977-982 (1996).
4. M. Thompson, L. Owen, K. Wilkinson, R. Wood, A. Damant. "A comparison of the Kjeldahl and Dumas methods for the determination of protein in foods, using data from a proficiency test", *Analyst* **127**, 1666-1668 (2002).
5. B. D. Ripley and M. Thompson. "Regression techniques for the detection of analytical bias", *Analyst* **112**, 377-383 (1987).
6. Analytical Methods Committee. "Fitting a linear functional relationship to data with error on both variables", AMC Technical Brief No 10. <www.rsc.org/amc/>.

APÊNDICE 6: COMO CONVÉM QUE OS PARTICIPANTES RESPONDAM AOS RESULTADOS DE ENSAIOS DE PROFICIÊNCIA

A6.1 Introdução

Participar de um programa de ensaios de proficiência se torna inútil se o participante não tiver proveito integral dos resultados de cada rodada, embora tomando cuidado ao interpretá-los. Os ensaios de proficiência representam, principalmente, uma ferramenta educativa que permite aos participantes detectar fontes inesperadas de erros em seus resultados.

Os ensaios, porém, não foram concebidos para serem ferramentas de diagnóstico. Consequentemente, eles apenas são úteis para o participante que já utiliza métodos validados e tem um sistema de CQ interno sob operação rotineira. Em tais condições, um resultado inesperadamente insatisfatório num ensaio de proficiência indica uma inadequação ou no método de validação ou no CQ interno ou, ainda mais comum, em ambos. (Um sistema de CQ interno adequado normalmente acusaria um problema na análise bem antes que o índice do ensaio de proficiência estivesse disponível. Há uma conexão demonstrável entre o desempenho dos participantes de um ensaio de proficiência e a eficácia do sistema de CQ interno em uso [1]).

Evitar interpretações equivocadas é particularmente importante quando o uso dos escores de ensaio de proficiência vai além do puramente científico, sendo usado, por exemplo, para fins de acreditação ou na literatura promocional de um laboratório. O participante deve levar em consideração a natureza estatística dos índices-z (*z-scores*), ao interpretá-los.

As seguintes orientações podem auxiliar os participantes na interpretação e uso adequados dos índices-z. (*z-scores*) Elas são reproduzidas mais ou menos intactas a partir de uma Súmula Técnica da AMC, com o consentimento da *Royal Society of Chemistry* [2].

A6.2 Ensaios de proficiência e acreditação

Os ensaios de proficiência são tão eficazes na detecção de problemas inesperados no trabalho analítico, que a participação em um programa (onde disponível) é geralmente considerada um pré-requisito para a acreditação do trabalho analítico. Avaliadores de acreditação terão a expectativa de encontrar um sistema documentado de respostas adequadas para quaisquer resultados que demonstrem exatidão insuficiente.

Convém que tal sistema inclua o seguinte:

- a definição dos critérios adequados para induzir as ações investigativas e/ou corretivas;
- a definição dos procedimentos investigativos e corretivos a serem usados, e um programa para sua implementação;
- o registro dos resultados de ensaios e conclusões, acumulados durante tais investigações; e
- o registro dos resultados subseqüentes mostrando medidas corretivas que tenham sido eficazes.

Esta seção provê orientação para habilitar químicos analistas a atender estas necessidades e demonstrar que tais necessidades foram atendidas.

A6.3 Procedimentos e documentação

Convém que os participantes implementem um procedimento documentado para investigar e tratar com índices-z (*z-scores*) insatisfatórios. A melhor maneira de faze-lo dependerá de como o programa de ensaios de proficiência é organizado. O sistema poderia tomar a forma de um fluxograma ou uma árvore de decisões, baseados nas considerações abaixo assim como nas necessidades particulares do participante. Entretanto, convém que o espaço para o exercício do discernimento profissional seja explicitamente incluído neste procedimento.

Os programas de ensaios de proficiência química geralmente estabelecem um critério para adequação aos propósitos que é amplamente utilizável nos campos pertinentes de aplicação. Mesmo que tal critério de "adequação aos propósitos" seja estabelecido, ele pode ou não ser apropriado para o trabalho individual de um participante para um cliente em particular. Este fator precisa ser considerado quando um participante organiza um sistema formal de resposta aos escores obtidos em cada rodada do programa. As principais possibilidades são abordadas abaixo.

A6.4 Efeito do critério de escore

A6.4.1 O programa de ensaios de proficiência usa um critério de "adequação ao uso"

O exemplo mais simples é quando o programa provê um critério de adequação aos propósitos σ_p como uma incerteza padrão e o usa para calcular índices-z (*z-scores*). Neste caso é importante perceber que σ_p é antecipadamente determinado pelos organizadores do programa para descrever a sua idéia de adequação aos propósitos: Não depende absolutamente dos resultados obtidos pelos participantes. O valor de σ_p é determinado de forma que possa ser tratado como um desvio padrão. Portanto, se os resultados são não tendenciosos(*unbiased*), têm distribuição normal, e o desvio padrão σ rodada a rodada é igual a σ_p, então os índices-z (*z-scores*) serão distribuídos como $z \sim N(0;1)$, isto é, na média, 1 entre 20 dos índices-z (*z-scores*) se situarão fora da faixa de ±2 e apenas aproximadamente 3 entre 1000 se situam fora de ±3.

Todavia, poucos (ou nenhum) laboratórios cumprem estes requisitos integralmente. Para resultados não tendenciosos, se o desvio padrão rodada a rodada σ de um participante for menor que σ_p, então menos pontos que os acima especificados estarão situados fora de seus respectivos limites. Se $\sigma > \sigma_p$, então uma grande proporção se situará fora dos limites. Na prática, a maioria dos participantes operam sob condições $\sigma < \sigma_p$, mas os resultados produzidos também incluem uma tendência de maior ou menor grandeza. Tais tendências freqüentemente comprometem a maior parte do erro total num resultado, e elas sempre levam ao aumento da proporção de resultados que se situam fora dos limites. Por exemplo, num laboratório

onde $\sigma = \sigma_p$, uma tendência de magnitude igual a σ_p aumentará por um fator de 8 a proporção de resultados que se situam fora de $\pm 3\sigma_p$.

Devido a tais conseqüências, é definitivamente útil registrar e interpretar índices-z (*z-scores*) para um tipo particular de análise na forma de um gráfico de Shewhart [3] ou outro gráfico de controle (ver também Apêndice 4). Se o desempenho de um participante fosse consistentemente adequado aos propósitos, um índice-z (*z-score*) fora da faixa de ±3 ocorreria muito raramente. Se isto acontecesse, seria mais razoável supor que o sistema analítico teria produzido uma séria tendência, do que supor um erro aleatório muito incomum. A ocorrência demonstraria que o laboratório precisaria implementar algum tipo de ação corretiva para eliminar o problema. Dois índices-z (*z-scores*) situando-se entre 2 e 3 (ou entre –2 e –3) poderiam ser interpretados da mesma maneira. Na verdade, todas as regras normais de interpretação dos gráficos de Shewhart poderiam ser empregadas, por exemplo, as regras Westguard [3] .

Além desse uso do gráfico Shewhart, pode valer a pena também testar os índices-z (*z-scores*) quanto à evidência de tendência de longo prazo, usando-se um gráfico CUSUM (Somas Acumuladas) ou um gráfico-J (Apêndice 4). Estes testes de tendência não são rigorosamente necessários: se os índices-z (*z-scores*) de um participante quase sempre preenchem os requisitos do critério de adequação aos propósitos, uma pequena tendência pode não ser importante. Entretanto, como se viu acima, qualquer grau de tendência pode levar a um aumento na proporção de resultados situados fora dos limites de ação e pode, portanto, valer a pena eliminar tal tendência. Convém que qualquer participante que decida ignorar este aspecto de tendência, o indique na especificação da ação investigatória. Em outras palavras, convém que o participante deixe claro para os avaliadores da acreditação que a decisão de ignorar tendências é deliberada e bem fundamentada, e não simples negligência.

A6.4.2 O programa de ensaios de proficiência não usa um critério de "adequação aos propósitos" adequado

Em alguns programas de ensaios de proficiência, o escore não é baseado na "adequação aos propósitos". O provedor do programa calcula o escore a partir dos resultados dos participantes apenas (i,é., sem referência externa a requisitos reais). Mais freqüentemente, um participante pode achar que o critério de adequação aos propósitos usado pelo provedor do programa é inadequado para certas classes de trabalho implementadas pelo laboratório. Participantes individuais em tais programas podem precisar calcular os seus próprios escores com base na adequação aos propósitos. Isto pode ser obtido de uma maneira direta, através dos métodos descritos abaixo.

Convém que o participante acorde com o cliente um critério de adequação aos propósitos específico na

forma de uma incerteza padrão σ_{ffp}, e a use para calcular o índice-z (*z-score*) modificado dado por

$$z_L = (x - x_a)/\sigma_{ffp}$$

a fim de substituir o índice-z (*z-score*) convencional (ver Seção 3.5.4). Convém que o valor designado x_a seja obtido do próprio programa. Se houver vários clientes com diferentes requisitos de exatidão, pode haver vários escores para cada resultado. Estes resultados poderiam ser tratados exatamente da mesma maneira acima recomendada para índices-z (*z-scores*), isto é, com os tipos usuais de gráficos de controle. Como o valor designado do analito não é conhecido pelo participante no momento da análise, um critério de adequação aos propósitos tem que ser geralmente especificado como uma função de *c*, a concentração do analito, conforme mostra a Seção 3.5.

A6.5 Como investigar um índices-z (*z-scores*) insatisfatório

A investigação de um índice-z (*z-score*) insatisfatório está intimamente ligada ao CQ interno [3]. Em circunstâncias normais, um participante de um ensaio de proficiência fica sabendo sobre um índice-z (*z-score*) insatisfatório dias ou semanas depois que a corrida de análise aconteceu. Em análise cotidiana, no entanto, convém que qualquer problema substancial que afete toda uma corrida seja detectado prontamente pelos procedimentos de CQ interno. A causa do problema seria corrigida imediatamente. A corrida contendo o material de ensaio de proficiência teria então que ser re-analisada, e um resultado presumivelmente mais exato apresentado ao programa de ensaio de proficiência. Assim , um índice-z (*z-score*) insatisfatório *inesperado* mostra ou que (a) o CQ interno é inadequado, ou que (b) o material de ensaio de proficiência, separadamente dos materiais de ensaio na corrida analítica, foi afetado por um problema. Convém que os participantes considerem ambas possibilidades.

A6.5.1 Falhas em sistemas de CQ interno

Uma falha comum de CQ interno acontece quando o material de CQ interno é inadequado para o material de ensaio típico. Convém que um material de CQ interno seja o mais representativo possível de um material de ensaio típico, no que diz respeito à matriz, separação, especiação, e concentração do analito. Apenas assim o comportamento do material de CQ interno pode ser um guia útil para toda a corrida. Se os materiais de ensaio diferem grandemente em algum destes aspectos, o uso de mais de um material de CQ interno pode ser benéfico. Por exemplo, se a concentração do analito varia consideravelmente entre os materiais de ensaio (digamos, acima de duas ordens de magnitude), convém que dois materiais de CQ interno diferentes sejam considerados, com concentrações perto dos extremos da faixa usual. É especialmente importante evitar usar uma simples solução padrão do analito como um material de CQ interno substituto para uma material de ensaio com uma matriz complexa.

Outro problema pode surgir se o sistema de CQ interno está focado apenas na precisão entre corridas e é negligente quanto a tendências no resultado médio. Tais tendências podem levar a um problema, seja ou não seja o material de CQ interno adequado ao tipo usual de material de ensaio (e, por conseqüência, ao material de ensaio de proficiência). Por conseqüência, é importante comparar o resultado médio com a melhor estimativa possível do valor verdadeiro para o material de CQ interno. A obtenção de tal estimativa requer rastreabilidade externa ao laboratório principal. Rastreabilidade externa poderia ser obtida, por exemplo, através de referência a MRCs de matriz comparável, ou submetendo-se o material de CQ interno candidato a um estudo interlaboratorial de algum tipo.

A6.5.2 Um problema com o material de ensaio de proficiência

Se o participante está convencido de que o sistema de CQ interno pode ser demonstrado como isento de tendências, deve-se suspeitar de um problema exclusivo do material de ensaio de proficiência. O resultado insatisfatório pode ser conseqüência de um erro relacionado ao manuseio do material de ensaio de proficiência (por exemplo, o registro de um peso ou volume incorreto). Convém que isso seja verificado. Alternativamente, um tipo inesperado de tendência (tal como uma interferência não previamente notada ou uma recuperação extraordinariamente baixa) poderia ter afetado exclusivamente o material de ensaio de proficiência ou o processo de medição. Uma conclusão válida a esta altura poderia ser que o material de ensaio de proficiência é suficientemente diferente do material de ensaio típico para tornar o índice-z (*z-score*) inaplicável à tarefa analítica que está sendo implementada.

A6.5.3 Testes diagnósticos

Um índice-z (*z-score*) insatisfatório indica um problema, mas não oferece um diagnóstico, de forma que um participante geralmente precisa de mais informações para determinar a origem de um resultado insatisfatório. Como primeiro passo, convém que o participante reexamine os registros da corrida de análise que utilizou o material de ensaio de proficiência. Convém que se procure pelos seguintes aspectos:

- cálculos com erros sistemáticos ou esporádicos;
- uso de pesos ou volumes incorretos;
- indicações "fora de controle" dos gráficos rotineiros do CQ interno;
- brancos extraordinariamente altos; e
- recuperações insatisfatórias, etc.

Se estas ações não resultarem em esclarecimento, então medições adicionais são necessárias. A ação óbvia é re-analisar o material de ensaio de proficiência em questão na próxima corrida de análise de rotina. Se o problema desaparecer (isto é, o novo resultado levar a um índice-z – *z-score* - aceitável), o participante pode ter que atribuir o problema original a um evento esporádico de causa desconhecida. Se o resultado insatisfatório persistir, será necessária uma investigação mais aprofundada. Esta poderia ser implementada através da análise de uma corrida contendo materiais de ensaio de proficiência de rodadas anteriores do programa e/ou MRCs apropriados, caso disponíveis.

Se o resultado insatisfatório ainda é obtido para o material de ensaio de proficiência sob investigação, mas não aparece nos resultados para os outros materiais de ensaio de proficiência e MRCs, então é provável que se origine de uma propriedade exclusiva do material, possivelmente uma interferência ou efeito matriz inesperado. Tal. descoberta pode demandar estudos mais extensivos a fim de se identificar a causa da interferência. Além disso, o participante pode precisar modificar o procedimento analítico de rotina a fim de acomodar a presença do
interferente em futuros materiais de ensaio. (Contudo, ele pode saber que os materiais de ensaios de rotina nunca conteriam tal interferente e assim decidir que o índice-z (*z-score*) desfavorável não é aplicável àquele laboratório em particular).

Se o problema é generalizado entre os resultados dos antigos materiais de ensaios de proficiência e MRCs, há um provável defeito no procedimento analítico e um defeito correspondente no sistema de CQ interno. Ambos demandariam atenção.

A6.5.4 Informações extras de resultados multianalitos

Alguns ensaios de proficiência envolvem métodos, tais como a espectrofotometria de emissão atômica, que podem determinar simultaneamente uma variedade de analitos a partir de uma única amostra de ensaio e um único tratamento químico. (Métodos cromatográficos que determinam uma variedade de analitos em rápida sucessão pode também ser considerados como "simultâneos" na presente discussão). Informação adicional que seja diagnóstica pode algumas vezes ser recuperada de resultados multianalitos de um material de ensaio de proficiência. Se todos ou a maioria dos analitos tem resultados insatisfatórios e são afetados aproximadamente no mesmo grau, a falha deve estar numa ação que afeta todo o procedimento, tal como um erro na pesagem da porção de ensaio ou na adição de um padrão interno. Se apenas um analito é adversamente afetado, o problema deve estar na calibração para aquele analito ou num aspecto exclusivo da química daquele analito. Se um subconjunto substancial de analitos é afetado, os mesmos fatores se aplicam. Por exemplo, na análise elementar de rochas, se um grupo de elementos dá baixos resultados, poderia ser produtivo verificar se o efeito seria rastreável à dissolução incompleta de uma das fases minerais componentes da rocha, na qual aqueles elementos estão concentrados. Alternativamente, poderia haver uma mudança espectroquímica causada pela variação na operação do sistema nebulizador, ou do próprio plasma, que afeta alguns elementos mais do que outros.

A6.5.5 Um valor designado com suspeita de ser tendencioso

A maioria dos programas de ensaio de proficiência usa um consenso de participantes como valor designado. Há raras alternativas práticas. Contudo, o uso do consenso traz a possibilidade de que haja, entre um grupo de laboratórios usando principalmente um método analítico tendencioso, uma pequena minoria de participantes que usa um método livre de tendências. Este subconjunto minoritário pode produzir resultados que se desviam do consenso e geram índices-z (*z-scores*) "inaceitáveis". Na prática, tal ocorrência não é comum, mas também não é desconhecida, particularmente quando novos analitos ou materiais de ensaio estão sendo submetidos a ensaios de proficiência. Por exemplo, a maioria dos participantes poderia estar usando um método com propensão a uma interferência não detectada, enquanto que uma minoria detectou tal

interferência e desenvolveu um método que supera tal problema.

Freqüentemente o problema é imediatamente aparente para os participantes afetados, porque eles usaram um método que é baseado num entendimento mais profundo dos procedimentos químicos do que aquele usado pela maioria dos participantes. Mas o problema não é visível para os outros participantes ou para o provedor do programa. Se um participante suspeita que se encontra nesta posição, o curso correto de ação, após passar por todos os passos descritos acima, é enviar ao provedor dos ensaios de proficiência detalhes acumulados da evidência de que o valor designado é defeituoso. O provedor normalmente terá acesso aos registros dos métodos usados pelos outros participantes e poderá estar na posição de constatar a reclamação imediatamente. Alternativamente, o provedor pode por em ação uma investigação do problema num prazo mais longo, a qual resolveria a discrepância no devido tempo.

REFERÊNCIAS

1. M. Thompson and P. J. Lowthian. "Effectiveness of analytical quality control is related to the subsequent performance of laboratories in proficiency tests", *Analyst* **118**, 1495-1500 (1993).
2. Analytical Methods Committee. "Understanding and acting on scores obtained in proficiency testing schemes", AMC Technical Brief No 11. <www.rsc.org/amc/>.
3. M. Thompson and R. Wood. "Harmonised guidelines for internal quality control in analytical chemistry laboratories", *Pure Appl. Chem.* **67**, 649-666 (1995).

APÊNDICE 7: GUIA DOS ENSAIOS DE PROFICIÊNCIA PARA USUÁRIOS FINAIS DE DADOS

As perguntas e respostas abaixo são baseadas em dúvidas relatadas por usuários finais de dados analíticos. Convém que a interpretação de resultados de ensaios de proficiência em química analítica seja conduzida com a colaboração de um químico analítico.

O que são ensaios de proficiência?

Os ensaios de proficiência consistem em um sistema interlaboratorial para testar regularmente a exatidão que os laboratórios participantes podem alcançar. Em sua forma usual, os organizadores do programa distribuem porções de um material homogêneo para cada um dos participantes, os quais o analisam sob condições rotineiras e relatam os resultados para os organizadores. Os organizadores compilam os resultados e informam aos participantes o resultado final, geralmente na forma de um índice relacionado à exatidão do resultado.

Qual a diferença entre ensaios de proficiência e acreditação?

Os organismos de acreditação requerem que os laboratórios analíticos participem de um programa apropriado de ensaios de proficiência onde disponível, e demonstrem um sistema para tratamento dos resultados. Este é apenas um dos muitos requisitos da acreditação.

Que tipos de materiais são distribuídos?

Os materiais distribuídos são os mais similares possíveis aos materiais que são habitualmente analisados, de maneira que os resultados do programa representem a capacidade dos laboratórios ao trabalhar sob condições rotineiras.

Para que servem os ensaios de proficiência?

O principal propósito dos ensaios de proficiência é auxiliar os laboratórios na detecção e correção de qualquer resultado de exatidão não aceitável dos resultados relatados. Em outras palavras, é concebido como um sistema educativo para informar aos participantes se eles precisam modificar seus procedimentos. Os ensaios de proficiência não foram concebidos para qualquer outro propósito, embora seus resultados, com a devida consideração de suas limitações, possam ser usados e combinados com outras informações para outros propósitos específicos.

Por que há resultados analíticos não exatos?

Todas as medições dão margem a não exatidão, tecnicamente conhecidas como “erros” na comunidade metrológica. (Aqui a palavra “erro” não significa que foi feito algo errado, mas simplesmente que o resultado de um processo de medição varia.). Erros surgem por causa da variação inevitável no procedimento físico ou químico empregado para se fazer a medição. Medições da concentração química requerem procedimentos muito mais complicados que as medições físicas típicas, tais como comprimento ou tempo. A medição de um comprimento a uma exatidão de uma parte em um milhão é um procedimento direto, contudo as medições químicas raramente podem ser feitas com uma exatidão melhor do que uma parte em cem. Na maioria das vezes, a exatidão nem é tão boa quanto essa, especialmente se as concentrações são muito baixas, por exemplo, quando se determina resíduos de pesticidas em alimentos.

A exatidão disponível é boa o suficiente?

Tudo depende da aplicação. Algumas análises têm que ser rigorosamente exatas. Por exemplo, na determinação do valor comercial de um carregamento consignado de sucata de ouro, o conteúdo de ouro tem que ser determinado com o menor erro possível (menos de uma parte em mil) porque um pequeno erro poderia ser equivalente a muitos milhares de Euros. Em outras aplicações, por exemplo, na determinação de cobre no solo, uma exatidão de uma parte em dez provavelmente seria suficiente: não importa se o valor verdadeiro é 20 ou 22 ppm, uma vez que a única decisão a tomar é se o nível está acima ou abaixo de 200 ppm. O custo também é considerado. Como uma regra geral, aumentar a exatidão de uma medição por um fator de 2 diminui a chance de uma decisão incorreta (isto é, dispendiosa), mas aumenta os custos de análise por um fator de 4. Estas considerações são conhecidas como “adequação ao uso”.

Como os programas de ensaios de proficiência avaliam individualmente a exatidão de laboratórios?

A maioria dos programas converte o resultado do participante num “índice-z (*z-score*)”. Este índice reflete duas características separadas, (a) a real exatidão alcançada (isto é, a diferença entre o resultado do participante e o valor aceito como verdadeiro), e (b) o julgamento do organizador do programa quanto a qual grau de exatidão que é “adequado aos propósitos”.

Como interpretar os índices-z (*z-scores*)?

Os índices-z (*z-scores*) devem ser interpretados com base em estatística (probabilística), e isto requer conhecimento especializado. Contudo, as regras simples abaixo se aplicam:

- Um índice de zero significa um resultado perfeito. Isto raramente acontecerá mesmo nos mais competentes laboratórios.
- Laboratórios que atendam o critério de "adequação ao uso" do programa de ensaios de proficiência comumente produzirão índices situados entre –2 e 2. Poder-se-ia esperar que eventualmente produzissem um valor ligeiramente fora desta faixa, aproximadamente 1 vez em 20, de maneira que um evento isolado deste tipo não tem tanta importância. O sinal (+ ou -) do índice indica um erro positivo ou negativo, respectivamente.
- Um índice fora da faixa de -3 a 3 seria muito incomum para um laboratório operando sob o critério dado de "adequação ao uso", e isto é indicativo de que o requisito de exatidão não foi atendido (pelo menos naquela ocasião). Convém que a causa do evento seja investigada e remediada.

Quais erros são comumente cometidos ao se usar os índices-z (*z-scores*)?

É importante não dar uma interpretação exagerada aos índices-z (*z-scores*). Isto pode acontecer de várias maneiras, tais como:

- A comparação de índices-z (*z-scores*) inter-rodadas ou interlaboratoriais tem que ser feita com grande cautela. Um laboratório individual, operando consistentemente alinhado com o critério de **"**adequação ao uso**"**, produziria tipicamente índices-z (*z-scores*) em rodadas sucessivas dentro da faixa de –2 a +2: o seguinte conjunto [0,6; - 0,8; 0,3; 1,7; 0,7; -0,1] seria típico. Os pequenos sobe-e-desce entre os índices não indicam uma mudança de desempenho: eles surgem por acaso. Assim, 1,7 não é "pior" que 0,3, e isto não indica deterioração do desempenho.
- Por causa desta "variação natural", não é válido formar um "grupo de iguais" de laboratórios, com base em seus índices-z (z-scores) numa rodada. Não é válido reivindicar que um laboratório com índice-z (*z-score*) de 0,3 numa dada rodada é melhor do que outro com índice-z (*z-score*) de 1,7.
- Julgamentos baseados em índices-z (z-scores) médios também requerem cautela. Não convém usar índices-z (z-scores) médios obtidos sobre uma variedade de analitos; eles podem muito bem mascarar o fato de que um dos analitos consistentemente dá índices-z (*z-scores*) insatisfatórios. Índices médios do mesmo analito através de várias rodadas podem ser mais úteis, mas ainda requerem a interpretação de um especialista.

Quais são as limitações dos ensaios de proficiência?

- Os ensaios de proficiência têm que ser realizados dentro do contexto de um sistema completo para a qualidade adequada, em cada laboratório. Não podem ser usados como um substituto para o CQ de rotina. Eles não representam, de forma isolada, um meio suficiente para validação de métodos analíticos, nem de treinamento de analistas individuais.
- Os ensaios de proficiência provêm ao laboratório participante apenas uma indicação de problemas, caso eles estejam presentes. Eles não provêm nenhum diagnóstico para auxiliar a resolver o problema.
- A aprovação num ensaio de proficiência para um analito não sugere que um laboratório é igualmente competente na determinação de um analito a ele não relacionado.